AD-A060 338 HUMAN RESOURCES RESEARCH ORGANIZATION ALEXANDRIA VA F/G 5/10
RESEARCH ON METHODS OF SYNTHETIC PERFORMANCE TESTING.(U)
APR 76 W C OSBORN, J P FORD DAHC19-74-C-0059
UNCLASSIFIED HUMRRO-FR-CD(L)-76-1 NL

1 OF 2
ADA
060338

E8800

ADA060338

HumRRO

LEVEL II

DDC FILE COPY

DDC
RECEIVED
OCT 25 1978
F

HUMAN RESOURCES RESEARCH ORGANIZATION

78 10 12 010

This document has been approved for public release and sale; its distribution is unlimited.

The Human Resources Research Organization (HumRRO) is a nonprofit corporation established in 1969 to conduct research in the field of training and education. It was established as a continuation of The George Washington University, Human Resources Research Office. HumRRO's general purpose is to improve human performance, particularly in organizational settings, through behavioral and social science research, development, and consultation.

Human Resources Research Organization
300 North Washington Street
Alexandria, Virginia 22314

ACCESSION for
NTIS White Section
DDC Buff Section
UNANNOUNCED
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
DIST. AVAIL. and/or SPECIAL
A

(1)

FINAL REPORT

RESEARCH ON METHODS OF SYNTHETIC PERFORMANCE TESTING

by

William C. Osborn and J. Patrick Ford

April 1976

Prepared for:

U.S. Army Research Institute for the Behavioral
and Social Sciences
1300 Wilson Boulevard
Arlington, Virginia 22209

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street
Alexandria, Virginia 22314

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: HumRRO-FR-CD(L)-76-1
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): 6 Research on Methods of Synthetic Performance Testing,
5. TYPE OF REPORT & PERIOD COVERED: 9 Final Report.
6. PERFORMING ORG. REPORT NUMBER: 14 HUMRRO-FR-CD(L)-76-1
7. AUTHOR(s): 10 William C. Osborn and J. Patrick Ford
8. CONTRACT OR GRANT NUMBER(s): 15 DAHC 19-74-C-0059
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Human Resources Research Organization, 300 North Washington Street, Alexandria, Virginia 22314
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 11 12 102p.
11. CONTROLLING OFFICE NAME AND ADDRESS: U.S. Army Research Institute for the Behavioral and Social Sciences; 1300 Wilson Boulevard, Arlington, Virginia 22209
12. REPORT DATE: April 1976
13. NUMBER OF PAGES: 90
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report):

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:

Research performed by HumRRO Central Division, Louisville Office, Louisville, Kentucky.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

synthetic performance tests
criterion testing
performance evaluation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Synthetic performance testing is based on concepts of simulation and task-element sampling. The objective is to develop efficient tests, tests that conserve administrative resources without sacrificing validity. This report describes the preliminary exploration of such an approach. First, job tasks were analyzed to identify testing problems that synthetic testing might solve. Next, a tentative model of synthetic test development was proposed. Finally, portions of the model were explored experimentally.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 405 260

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

SUMMARY

BACKGROUND

Efficient tests are the goal of the test developer and administrator. Two assumptions can be made concerning test types; first that a full performance test (demonstration of the actual criterion behavior in a realistic setting) is the most valid type of test: second, that a group administerable test yielding scorable task product and process information is the most feasible type of test. Any test, then, that is both full performance and group administerable -- valid and feasible -- is an efficient test and should present no problem for the developer or administrator. Experience suggests, however, that many job-tasks cannot be simply translated into efficient tests. The nature of these tasks, together with time and cost constraints, create problems for the test developer; problems which all too often are circumvented by resorting to substitute tests of questionable efficiency. A possible solution to these problems is based on the concepts of simulation and task element sampling and is termed synthetic performance testing.

A synthetic performance test is conceived of as a job performance test that has been degraded to some degree in the range of task elements covered or in the fidelity of stimulus/response features. The intent is to connote a process of synthesis by which the substructure of a job-task is used as the basis for selectively constructing alternate forms of a test; each test representing (at least theoretically) a more or less optimal blend of validity and feasibility.

PURPOSE

The concept of synthetic performance testing seems reasonable, but needs further development and empirical verification before it can be codified into procedures useful to test developers. The research reported here attempted to chart the areas of synthetic performance testing by focusing on three major objectives:

1. Identify testing problems that forestall the use of efficient tests.
2. Develop a tentative model of synthetic testing procedures.
3. Explore portions of the model experimentally.

ANALYSIS OF TESTING PROBLEMS

Features of job-tasks that are problematic in developing efficient performance tests were identified and classified. Two categories, task conditions and task behavior, were identified as potential sources of difficulty for the test developers in achieving fully relevant yet feasible tests. Testing problems within each category were then identified and reviewed from the standpoint of their implications for synthetic testing.

The testing problems are:

<u>Task Conditions</u>

1. Scarce
   - Equipment/Facility
   - Terrain
   - Personnel
2. Dangerous
3. Variable
   - Surround
   - Display
4. Latent

<u>Task Behavior</u>

1. Long process
2. Transient process
3. Affective

The potential relevance of synthetic methods to the problem areas was discussed. It was concluded that the problems of dangerous, scarce, and variable task conditions are potentially resolvable by synthetic test methods. The same was concluded for tasks involving a long process. On the other hand, problems of transient and affectively controlled task behavior, or tasks subject to latent conditions, appear less amenable to solution by synthetic methods.

PRELIMINARY MODEL FOR SYNTHETIC TESTING

A procedural model for synthetic testing was conceptualized. The model attempts to lead the test developer to an efficient method for testing a task by progressing through a series of decisions about a task's response characteristics. The simple taxonomy of task characteristics is based on the assumption that a task can be characterized in three dimensions: a behavioral component, a skill component, and an affective component. The model attempts to be exhaustive by covering all problems identified as deterrents to efficient testing. Captured in the model are the three most problematic and interactive aspects of the task-test domain: type of task, task element sampling, and test method. Moreover, the general framework seems useful as a way of codifying guidelines for the development of synthetic tests.

PILOT EXPERIMENTATION

Two parts of the preliminary model were selected for pilot experimentation:

1. Knowledge testing of low-skill psychomotor tasks.
2. Synthetic testing of skilled psychomotor tasks.

Each experiment is summarized below:

Low-Skilled Procedural Tasks are manual tasks relatively unrestrained by time and composed of discrete steps, each of which can be performed with one or two guided practice trials. The purpose of this experiment was to assess four methods of knowledge testing in terms of their relative and absolute correlation with "hands-on" task proficiency for high and low mental ability Ss.

A hands-on performance test and four versions of a knowledge test were developed for each of three job tasks. One test version was a conventional multiple-choice test; the other three employed pictures but used different methods of eliciting task knowledge.

A number of interesting though tentative findings emerged:

1. High correlations between task knowledge and task performance strongly supported the hypothesis that performance on low-skill procedural tasks is mediated by knowledge.

2. Substantial differences among the methods of knowledge testing were not found.

3. Correlations between knowledge and performance were not significantly different for low versus high mental ability Ss, though the latter group had higher average scores on the knowledge tests.

Skilled motor tasks present a special challenge to synthetic testing. Three small demonstration studies were designed to explore two assumptions not reflected in the preliminary model. The assumptions are:

1. In any motor skill there is at least one key response feature that distinguishes masters from nonmasters.

2. In eliciting even a facsimile of that feature through part-task simulation, the response of a master will differ generically from that of a nonmaster.

Study I attempted to demonstrate the effectiveness of a synthetic test method in which aspects of the feedback complex were degraded. The task was aim-firing the .45 caliber pistol. The results indicate that it may well be possible to test this type of skill validly with substantial reduction in external feedback fidelity.

Study II was designed to determine whether a pattern of response or "signature" distinguishing masters from nonmasters could be identified for a motor skill. The relevant task again was pistol firing, and the response feature of interest was barrel movement in aiming. Unfortunately, apparatus problems prevented completion of this study.

Study III was a brief test to explore the possibility that cognitively mediated aspects of a motor skill could be identified, measured, and validated against performance. The task was typing. The results suggest the possibility that skilled task performance (mastery) may actually inhibit performance on a synthetic test of that task.

CONCLUSIONS

The results of the pilot experiments indicate that the model for synthetic performance testing is a promising one. The rationale and definitions underlying each of the decision points in the model need further development, and the method categories need expanding into

hypothesized procedures. Yet this conceptual work can go only so far without additional empirical data. Maturity of a synthetic approach to test development can be brought about best by the continued interplay of conceptual and experimental effort.

PREFACE

This is the final report on a project entitled, "Research on Methods of Synthetic Performance Testing." The project was directed toward the exploration of test methods and media comprising efficient substitute tests for use in those instances where full performance testing is not considered feasible. The report covers analysis of testing problems that preclude the use of efficient tests, a tentative procedural model for development of synthetic tests, and results of pilot experimentation on synthetic testing concepts.

Work reported here was conducted by the Human Resources Research Organization (HumRRO) under Contract No. DAHC 19-74-C-0059 with the U.S. Army Research Institute for the Behavioral and Social Sciences. The research was performed at HumRRO's Central Division (Louisville) under the supervision of William Osborn, who is Director of the Louisville Office and Project Director. Dr. Wallace W. Prophet is Director of the HumRRO Central Division. The project staff included J. Patrick Ford, Peter B. Wylie, Robert Vineberg, Roy C. Campbell, James H. Harris, and Jack R. Reeves. Mr. Eugene Johnson, the Contracting Officer's Technical Representative, provided administrative and technical guidance throughout the project and contributed substantially to the conceptual work reported here. LTC Willis G. Pratt, Military Chief of ARI's Fort Knox Field Unit, secured the soldiers who served as subjects for data collection.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

RESEARCH ON METHODS OF SYNTHETIC PERFORMANCE TESTING

INTRODUCTION

A job performance test may be defined as the controlled observation of job behavior under realistic and standardized job conditions. The development and use of such tests would seem to be straightforward: the job-relevant conditions for task performance are created and an acceptable criterion of performance defined. Then the trainee or job incumbent is asked to perform, and his performance is evaluated against the established criterion. Unfortunately, the nature of certain types of job-tasks, together with time and cost constraints, often create problems for the test developer. In circumventing these problems he frequently resorts to simplistic test procedures of questionable reliability or validity. More grave, however, is the fact that such compromises so often occur -- apparently either because of inadequate regard for the price one pays in diminishing reliability and validity, or because of a lack of awareness of alternate approaches. The most typical substitute for performance testing is the written multiple-choice test of job knowledge.

It is worthwhile to attempt clarification of the term "performance test" in relation to what appears to be a separate and distinct class of tests descriptively termed "paper-and-pencil test," or "knowledge test," or simply "written test." These kinds of labels reflect artificial distinctions and are misleading. They normally refer to the ubiquitous multiple-choice knowledge test presented in a verbal medium. Yet, a true performance test for many clerical tasks

would also be "paper-and-pencil"; assessing the performance of one who operates an information center would involve essentially "knowledge testing"; and, performance measures for jobs involving journalistic or written communication tasks would lead to the use of "written tests." It is even possible for a job performance to consist of a true multiple-choice response, as with a surgical assistant selecting on command the proper instrument from an array; or a tank gunner engaging the most threatening target when confronted with multiple enemy emplacements.

The one really important, though implicit, feature of the so-called written (or knowledge) class of tests is that they are administrable to large numbers of people at once. The reason such tests are group-administrable is two-fold: a) they do not involve scarce equipment or space which often is either unavailable in quantity or too costly to tie up for group testing purposes, and, b) they yield a recorded product(s) which may be collected by the tester and scored later -- both in terms of accuracy of the product and reasons for any inaccuracies or failures.

So, to achieve some perspective on types of tests it might be well to begin by casting out most of the casual or superficially descriptive labels, and make two assumptions: first, that a full performance test (demonstration of the actual criterion behavior in a realistic criterion setting) is the most <u>valid</u> type of achievement test; second, that a group administerable test yielding scorable task product and process information is the most <u>feasible</u> type of achievement test. Any test, then, that is both full performance

and group administerable -- valid and feasible -- is an efficient test, and should present no problem to the developer or administrator. On the other hand, experience suggests that many tasks cannot be simply translated into efficient tests. When this occurs some type of substitute test is created -- hopefully one that minimizes loss in validity and feasibility.

This project was directed toward the exploration of test methods and media that lead to highly efficient substitute tests. A synthetic test, as tentatively conceived, is a job performance test that has been degraded to some degree in the range of task elements covered or in the fidelity of stimulus/response features. The term is used to indicate a somewhat broader range of possible alternative approaches to testing than implied in related terms such as "simulated" or "symbolic" test. The intention is to connote a process of synthesis by which the substructure of a job task is used as the basis for selectively constructing alternate forms of a test, each representing (at least theoretically) a more or less optimal blend of validity and feasibility. In some cases this may be achieved through simulation; that is, by substituting for stimuli in either the task display or the surround, or by requiring a substitute response. In other cases, efficient tests may be created by testing on a subset of task elements, regardless of whether simulation is used. Thus, synthetically generated alternatives to fully relevant performance tests may vary in two major dimensions, fidelity and scope.

Examples can be found (Glazer, 1954; Frederiksen, 1957; Osborn, 1970) of what is here termed synthetic tests. Unfortunately, they represent little more than examples, as they deal with substitute test methods limited to specific types of tasks. None reflects an underlying systematic basis for generalizing the approach to other types of tasks - or, for that matter, even to other similar tasks. What is needed, before work proceeds toward the development of synthetic test procedures, is an analysis of the types of job tasks for which synthetic tests offer potentially efficient means of performance evaluation. Even more basically, we should first answer the questions, "Why can't the performance of _any_ task be evaluated with an efficient test?"; "What are the traits of certain tasks that prevent efficient testing?" Answers to these questions should enable us to identify and classify the kinds of performance testing problems which can logically be addressed via synthetic testing methods. Methods can then be conceptualized for combinations of tasks and testing problems, and evaluated empirically.

The work reported here is an attempt to chart the area of substitute or synthetic performance testing. The report covers three phases of research. First, job tasks were analyzed to identify testing problems that forestall the use of fully efficient tests, and to estimate the usefulness of synthetic testing approaches in circumventing these problems. Next, a tentative model of synthetic testing procedures was developed. And, finally, portions of the model were explored experimentally. Results of this work are reported in the following sections.

ANALYSIS OF TESTING PROBLEMS

The purpose of this phase of the research was to identify and classify features of job tasks that are problematic in developing efficient performance tests. An inductive approach was used in which many job-tasks were analyzed in terms of difficulties presented to the test developer. The difficulties were organized into classes of testing problems and keyed to major task dimensions. Problem classes were then reviewed from the standpoint of their implications for synthetic testing.

METHOD

A sample of 100 tasks was selected from an existing inventory of approximately 700 tasks spanning all skill levels for an Army job (Reconnaissance Specialist). Tasks were selected proportionally and randomly from each of 31 content areas included in this inventory. The most heavily represented content areas in the initial sample were: Machineguns (12 tasks), Communication (6 tasks), Leadership (6 tasks), Tracked Vehicles (6 tasks), Tactics (6 tasks), First Aid (5 tasks), and Land Navigation (5 tasks).

As an additional precaution to insure a reasonable variety of job tasks, the initial sample was subjected to another screening. This was accomplished using a simple taxonomic structure defined by three levels of task behavior and three classes of task display. The levels of task behavior were taken from Ammerman and Melching (1966) and are characterized as follows:

- Specific Task - A particular work activity, with a clear beginning and ending point, that is performed under a specific set of task conditions.

- Generalized Skill - A relatively specific activity performed under similar but not identical task conditions.

- Generalized Behavior - A manner of behavior or way of doing things; e.g., application of values and principles.

The types of task display - People, Data, Things - were taken from the Department of Labor's occupational analysis work (1965). Combining these two factors produced nine categories into which tasks were sorted.

As expected, classification of the 100 tasks resulted in a high concentration of Specific Tasks and Generalized Skills, chiefly dealing with Things. Virtually absent were Generalized Behaviors. While the distribution accurately profiled the 11D's job, it was not considered satisfactory for our purposes. Therefore, the 600 remaining tasks were again screened specifically for Generalized Behaviors, as well as additional People and Data oriented tasks. The attempt was to over-sample in all categories but Specific Tasks and Generalized Skills pertaining to Things. The final set of 100 tasks was distributed as shown below:

		Task Display			
		People	Data	Things	
Task Behavior	Specific Task	4	11	31	46
	Generalized Skill	11	15	20	46
	Generalized Behavior	2	3	3	8
		17	29	54	100

Some representation was provided in each category, which was about all that could be hoped for given the nature of the Reconnaissance Specialists' job. Though less than ideal, the variety of tasks was considered adequate for the purpose of identifying the full range of potential testing problems.

Once the working set of 100 tasks was decided on, a fully relevant job performance test was conceptualized for each. In conceptualizing these tests, every effort was made to capture realistic aspects of the job situation in which the tasks would be performed. For example, in the task "Apply first aid measures for a fracture, sprain, or dislocation," consideration was given to (a) the types and numbers of casualties to which each trainee would be exposed; (b) the stressful environment in which the trainee would be required to perform; (c) the various methods by which the trainees' performance would be scored; and, (d) the time required to test each trainee. With a concept for the fully relevant test in mind, it was then evaluated from the standpoint of feasibility. Feasibility was judged principally in terms of the test's conformity to the following hypothetical but not unrealistic constraints:

- 50 men to be tested
- one hour to complete testing
- three test administrators or monitors
- no more than one item of major equipment per ten people
- go/no go and process scores required

In addition to problems created by the above constraints, other special features were noted which, if not dealt with by other than a conventional test mode, would severely jeopardize test validity or reliability.

With the assistance of a military subject matter specialist each of the 100 job tasks was explored in this manner, and the problems recorded by task in a narrative fashion. The full list of problems was then reduced by collapsing similar problems into categories. Finally, the problem categories were organized and redefined in terms of generalized task characteristics or dimensions.

RESULTS

Two major dimensions of job-tasks, Task Conditions and Task Behavior, were identified as potential sources of difficulty for the test developer in achieving fully relevant yet feasible tests. Outlined below is the structure of task characteristics which inhibit the development of performance tests, and which must be dealt with through simulation, task element sampling, or other means if near optimal compromises betwen validity and feasibility are to be achieved. The type and relative frequency of these testing problems is summarized in Table 1. A complete tabulation by task is given in Appendix A.

TABLE 1

FREQUENCY AND TYPE OF TESTING PROBLEMS
ANTICIPATED IN A SAMPLE OF 100 JOB TASKS

Problem		Frequency	Sample Task
Task Conditions			
Scarce		84	
Equipment/Facility	62		Zero 551 gun launcher
Terrain.	43		Navigate with a compass
Personnel.	34		Lead a security patrol
Dangerous.		33	Prepare a shaped charge
Variable		38	
Surround	5		Engage target under limited visibility
Display.	34		Camouflage a weapon
Latent		12	Give CBR alarm
Task Behavior			
Long Process		18	Conduct marksmanship training
Transient Process. . . .		43	Communicate by radio
Affective.		12	Maintain light discipline

Task Conditions

Scarce. A very large percentage of tasks involve task-relevant conditions (equipment, terrain or support personnel) that are costly or otherwise difficult to obtain in the quantity normally required to efficiently test a large number of personnel in a reasonable length of time. Of the sample of 100 tasks examined, 62 involved equipment (and facilities) which would normally not be available in sufficient quantity for any type of group testing. Because of terrain requirements full-field testing on 43 of the tasks would be difficult or impossible to carry out on other than an individual basis. Similar testing limitations would exist for 34 of the tasks which involve either (a) large numbers of specially trained personnel to participate as task-relevant conditions, or (b) support personnel to act as test controllers. A total of 84 tasks require one or more of these three types of scarce resources.

Dangerous. Thirty-three tasks were identified as having conditions which, if realistically created, would be either physically or psychologically hazardous to people being tested. In nearly all of these cases the conditions are sufficiently dangerous as to unequivocally preclude testing with full realism.

Variable. Standardization of conditions is fundamental to valid testing, yet 38 of the tasks present some problem with respect to maintaining control over task conditions. Five tasks involve aspect of the surround (e.g., wind and visibility) which would be most difficult to reproduce from one test session to another. In 34 tasks,

elements of the display presented control problems. Thirteen cases involved people as task-relevant aspects of the display, and the behavior of others would be difficult to standardize over test administrations. Other display conditions which would present standardization problems were identified in 21 tasks (e.g., materials available for camouflaging a weapon, or the condition of equipment to be serviced).

Latent. For lack of a more descriptive term, the word "latent" is used to characterize task relevant conditions which require detection as well as immediate reaction by the person being tested. Tasks with this feature may complicate the testing process in that a prolonged time period may be needed to accommodate the vigilance set required. Only 12 such tasks were identified including "Give CBR Alarm," "Perform duties of an interior guard," and, "Avoid poison plants." Testing Problems created by tasks of this type are similar to those encountered for affective task behavior as mentioned below.

Task Behavior

Lengthy Process. The time required to execute a task is an obvious factor to be considered in developing feasible tests. Eighteen tasks were judged to take more than an hour. In most of these cases the testing time was estimated at 1 to 4 hours. However, in the case of "Recommend personnel for promotion," full enactment of the task could well involve several days or even weeks if even minimally realistic conditions for personnel data collection were created.

Transient Process. In achieving efficient tests perhaps the most constraining task characteristic is that of a transient task process. The process in performing some tasks is preserved in or inferrable from the product or outcome of the task (e.g., "Prepare a written message," or "camouflage a weapon"). Tasks of this sort, unless prohibitively expensive equipment is involved, can be group tested. On the other hand, tasks in which process is not preserved in the product (e.g., "Communicate information over radio net" or "Ground guide a wheeled vehicle") must be administered on an individual basis in order for the tester to record performance. A variation of this problem occurs in some task products which, were it not for certain critical safety precautions that must be observed, otherwise preserve the process (e.g., splinting a broken leg or disassembling a weapon). Transient process was a characteristic of 43 of the tasks.

Affective. Twelve tasks were seen as having a substantial affective or "willingness to perform" component. This means that some unobtrusive method of testing would be called for. In testing on the task "Maintain light discipline," for example, the soldier must not know he is being tested if the test is to provide a valid measure of task performance. In the case of preventive maintenance tasks on weapons, deciding to perform the maintenance is at least as important a behavior as performing the maintenance itself. Realistic enactment of conditions for tasks of this sort presents problems from the standpoint of creatively and unobtrusively utilizing testing time to permit the behavior to be emitted.

IMPLICATIONS

Though little in the way of detailed implications for synthetic testing can be drawn at this point, certain general observations may be made. The potential relevance of synthetic methods to the problem areas is discussed below. Our judgment of the degree of relevance is reflected in the order in which the problems are discussed.

Dangerous Conditions. The area perhaps of highest priority for utilization of synthetic test methods is that in which the creation of realistic task conditions constitutes a danger to the man being tested. In these instances a fully relevant performance test is totally and unequivocally out of the question, and some type of simulation or part task testing is a necessity. This has long been recognized both as a training and testing problem. On one hand, such tasks are usually the most important to train and test effectively, for the very reason that they are dangerous; on the other, attempts to generate realistically a sense of danger may be viewed as unethical treatment of participants. Severe threat to life, limb or psychological stability is normally to be avoided in training and testing exercises. Yet simulation can offer a means of resolving this dilemma.

To be effective, a simulation does not have to duplicate or even approximate the actual stressful stimulus. It must provide a sufficient level of stress to instill a sense of psychological intensity or urgency on the part of the person being trained or tested, and its form should be such that it effectively duplicates the real stimulus in the control exerted over the criterion behavior.

A good example of this is seen in an innovative simulation used in an infantry platoon combat exercise (U.S. Army Research Institute, 1973). The principal feature of the method involves each man having a number on his helmet and an inexpensive scope mounted on his rifle; then, during the exercise a soldier may "kill" by correctly reporting an enemy's number, or "be killed" by allowing his number to be sighted by the enemy. Number size and scope power have been carefully calibrated from empirical data so that the probability of a simulated "kill" is highly correlated with the expected outcome in actual battle. The appeal of this simulation is that, without real bullets, the situational feedback exerts realistic control over players' behavior. Experienced soldiers report that mistakes leading to "death" in the simulated situation are virtually the same as those that cause people to be killed in combat.

Similar effects could be created for other dangerous tasks. An effective simulation in "preparing a shaped charge," for example, might involve a clay substitute for the plastic explosive with a buzzer assembly attached. If the sensitivity of the buzzer were closely calibrated to that of the actual explosive, handling of the mock explosive could be realistically tested (and trained).

The point here is that the simulation need not entail anything physically similar to the aversive stimulus, or even psychologically similar in intensity of its threat. The substitute stimulus threat must only enable a timely, accurate, unequivocal, and public record of response adequacy. The importance of carefully designed effective

simulations for tasks of this sort is apparent in light of the fact that the resulting test conditions constitute the most relevant criterion situation -- there neither is nor will be (short of a war environment) a better criterion against which to validate the simulation.

Scarce Conditions. Unavailable equipment, facilities or terrain is the traditional area for use of simulation. The fidelity of equipment simulations required for effective training has been the object of much study (e.g., Dougherty, et al, 1957; Cox, et al, 1965; Prophet and Boyd, 1970). A result of possible significance to synthetic test development is that relatively low level simulations are adequate for training simple procedural tasks, while high fidelity simulation is needed for highly skilled tasks. This generalization applies equally well to motor and perceptual skills. Though the latter may be less well researched, there is evidence for discrimination or recognition tasks that extremely high fidelity stimuli are necessary if training is to transfer positively to the field setting (Mackie, 1964; Baldwin, 1973).

The same notion may be tentatively generalized to problems of unavailable terrain. Where task behavior involves fine-grained perceptual discrimination pertaining to terrain, as in tactical driving or target identification, very high fidelity terrain simulation will likely be required. On the other hand, where task performance is relatively robust with respect to perceptual feedback from terrain

features - as in conducting an administrative movement, breeching a minefield, or possibly even in land navigation tasks - it would seem that fairly degraded terrain simulations could be used. There is evidence to suggest that, where miniaturization has been found appropriate in training, the degree of reduction and fidelity of terrain simulations are not highly significant factors (Baker, et al, 1964).

The problem of personnel as relevant conditions in task performance actually breaks down into two separate issues. First, there is the case in which people function as reasonably reliable, methodical and responsive elements of the environment, much like equipment or other inanimate task-relevant conditions; examples include, "conduct an administrative movement," "coordinate unit defense plans with adjacent units," "supervise a route reconnaissance." In such instances "personnel-objects" can probably be modeled through miniaturized mockup or other means without greatly reducing effective fidelity. The second class of task involving people as task-relevant conditions is that which requires skilled and adaptive interaction, whether verbal or physical, between the person being evaluated and the object person. As this second case centers around the problem of variable task conditions it will be discussed below.

Variable Conditions. Difficulties in controlling conditions for certain tasks presents a problem that certainly should be addressed by synthetic test methods. As with dangerous conditions - though to a lesser extent - variable conditions in some instances seriously inhibit good performance testing at most any cost. It is usually a

two-sided problem. If we attempt to control conditions by testing people individually in exactly the same setting on identical conditions, wear and tear from repeated administrations begin to leave tell-tale signs that prompt or mislead subsequent test subjects. On the other hand, it may not be possible to circumvent that problem by creating several identical sets of conditions. Conditions for "camouflaging a weapon" or "clearing fields of fire" simply cannot be duplicated many times over for realistic standardized testing. Similarly, creating a "standard" amount of dirt, rust, wear and operational deficiency in weapons so that men may be tested on their maintenance-services performance is simply not feasible. If in these cases we can assume that motor skill requirements for executing the task are minimal, and it is the perceptual and decision-making skills that are dominant, the effective use of synthetic substitutes is very possible.

The issue of people as variable task conditions was mentioned above. Tasks like "apply psychological first aid," "investigate complaints," "counter disruptive influences and acts," and "recommend personnel for promotion" require demonstration of interpersonal skills ranging from empathy to actual physical intervention. Here, the other person or group of people represent an important part of the environment to be controlled - that is, standardized - from one test administration to the next. People are difficult to standardize; and, other than through the use of well trained "standardized others" in a role playing mode, effective synthetic

solutions are not obvious.

Long Task Process. Where task performance is extremely time consuming, the concept of part-task testing would seem particularly useful. In navigation from point A to point B on the ground using map and compass, for example, a soldier might be tested on three elements: set an accurate compass heading, pace exactly ten meters, and calculate the number of paces necessary to arrive at point B. This would eliminate the time required to actually negotiate the full distance on foot, and presumably without significant loss in test validity. Breeching a minefield could be similarly tested by merely requiring the soldier to designate a route through the field and demonstrate his ability to probe for a mine.

Latent Conditions and Affective Behavior. Testing problems associated with tasks requiring an unalerted reaction to infrequent and unpredictable stimuli do not seem particularly amenable to solution by synthetic methods. The same holds for tasks with a substantial affective or "will do" (as opposed to "can do") component. In both cases effective test methods must center around creation of an unalerted and unobtrusive set - that is, the soldier must not be aware that he is being tested, at least unaware he is being tested on that particular task. Solutions to this problem lie in an approach that a) satellites the task on another task that is ostensibly the one being tested, or b) embeds the subject task in a job context of other tasks that are being tested as a functional module (Osborn, et al., 1974). This type of solution lies outside the domain of synthetic test methods.

Transient Task Behavior. Tasks in which the process is not preserved in task product or outcome also constitute a problem that is not directly relevant to synthetic testing. To avoid the requirement for a tester to observe and score each soldier's task performance as it is executed, all one has to do is record performance on audio (e.g., "communicate information over radio net") or video ("ground guide a wheeled vehicle") tape for scoring later. To the extent that competent test administrators are not available in quantity, or that other problems associated with individual testing prevail, the use of recordings may be more than justified. But again, as with the previous problem category, this type of solution does not constitute a synthetic test method.

The foregoing analysis of testing problems provided a basis for the next phase of the research, which was to develop preliminary guidelines for synthetic test development.

PRELIMINARY MODEL FOR SYNTHETIC TESTING

A variety of barriers to constructing efficient tests was identified in the first phase of research, and it became increasingly apparent that no single approach to synthetic test development would work for all tasks. A dilemma, not uncommon in new areas of research, surfaced at this point in the project: hypotheses about techniques of constructing efficient tests should be derived from theoretical or experimental work, yet little or no research has been done on the comparability of various test methods; on the other hand, good theoretical or experimental work depends on carefully formulated hypotheses. So, as a starting point, a procedural model was conceptualized. This model attempts to lead one to an efficient method for testing the task through a series of decisions about a task's response characteristics.

Several troublesome areas had to be resolved in the course of developing the model, even at a conceptual level. Since their resolution led to assumptions underlying the development, it is important that they be highlighted before presenting the model. The major issues pertained to task heterogeneity, test simulation and transfer effects, and stimulus-response dependence.

Task Heterogeneity. Job tasks must be treated generically in studies such as this, otherwise one is faced with the inefficiency of devising methods anew for each set of job tasks encountered. The eventual availability of engineering specifications for developing efficient tests depends largely on our success in identifying task

categories to which test methods can be tailored. Unfortunately, job tasks are designed to serve system requirements, not psychological taxonomies. They often consist of dissimilar sub-tasks or elements which collectively are difficult to slot in a task classification system. The problem is two-sided. On the one hand, it is difficult to place such a task in a single category of a classification scheme without one or more of the parts seemingly misclassified; on the other hand, if one partitions the task into elements and classifies each, the usefulness of the system suffers since the notion of task groups is lost. This problem was resolved, if not entirely eliminated, by suggesting that the most critical or difficult element of a task be identified, classified, and a test method suitable to that element be used in testing. "Critical" or "difficult element" is defined as that part of task performance which most often leads to failure -- information obtainable from performance test results or from the judgment of experienced job incumbents.

Task Simulation and Transfer. Simulation for training and testing have different standards of effectiveness. Effective simulation in training is evaluated in terms of minimizing time to learn the criterion task; whereas, effective simulation in testing is evaluated in terms of equivalence of performance, in a concurrent validity sense, under actual and simulated conditions. Moreover, there does not appear to be a unified set of principles which ties together what is known about transfer as a function of stimulus and response similarity (e.g., Osgood, 1953) and the effects of

overlearning the initial task. The joint effects of these factors are important to the design of effective test simulations. There is evidence indicating that, in their effects on transfer, interaction exists between task dissimilarity (initial versus criterion task) and the extent of training on the initial task (Weitz and Alder, 1973). This implies that there is an optimum point in initial learning at which a trainee should be transferred to the criterion task; otherwise, over or underlearning may produce poorer transfer. The implications for testing are less clear. But in attempting to apply this hypothesis to the design of test simulations, first we must reverse the paradigm; that is, think in terms of transfer from the criterion task to the simulated task. Secondly, we must recognize that job incumbents represent some distribution of skill, level of learning or mastery, on a criterion task, and our objective is to predict each person's skill level from his performance on a substitute or synthetic test of that task. Now, if overlearning (mastery) of a job task can interfere with performance on a synthetic test of that task, test media must be selected very carefully. Simply put, an effective simulation for training may not be effective for testing.

Stimulus-Response Interdependence. In considering a full range of simulation alternatives to tests of task performance, one is constrained by the interdependence of task stimulus and response. That is, it is not always sensible -- even if possible -- to vary stimulus and response fidelity independently of one another. The manipulative aspect of rifle disassembly, for example, cannot be

mediated with a picture of the weapon. Similarly, the cognitive aspect of a target selection task could be mediated by presentation of real targets (vehicles, troops, bunkers, etc.), but it probably would not be sensible to do so.

These issues, together with a task taxonomy and the barriers to efficient tests described earlier, were the basic considerations in developing the model.

THE MODEL

As shown in Figure 1, the model is designed to lead one to an efficient method of testing performance on a task, through a series of decisions about task characteristics. The underlying taxonomy of task characteristics is a simple one. It is based on the assumption that a task can be characterized in three dimensions:

- One of three categories of behavior -- cognitive, perceptual, psychomotor;
- One of two skill levels -- high (requires practice for mastery) or low (requires little or no practice for mastery);
- With or without a major affective or motivational component.

The model attempts to be exhaustive in the sense of covering all problems identified as deterrents to efficient testing. Thus, the affective or motivational dimension has been included even though its potential for resolution through synthetic testing was discounted. Beyond the following paragraph the report focuses exclusively on the left-hand branch of the model where skilled aspects of task behavior are charted into test methods.

One begins, as indicated at the top of the diagram, by determining

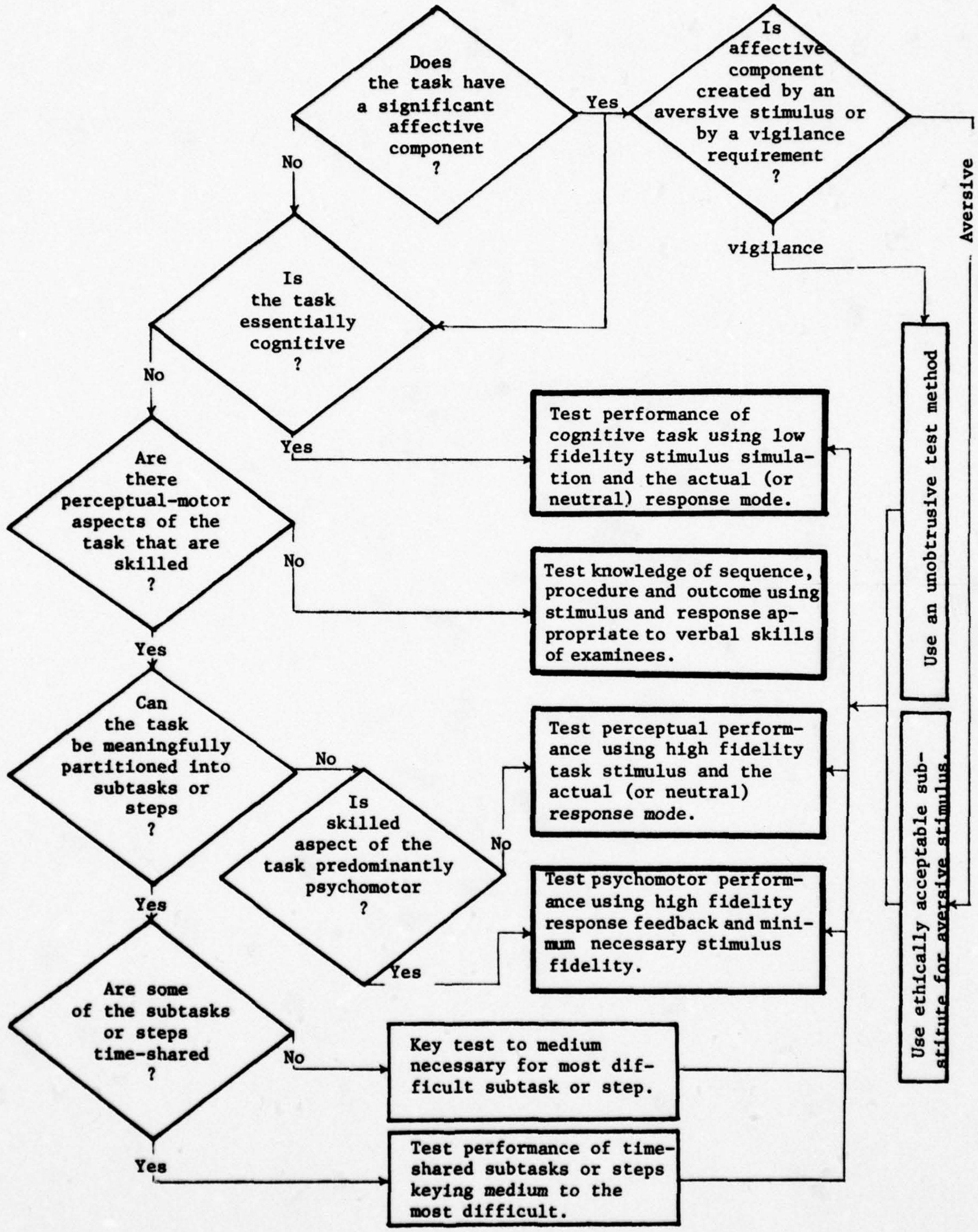


FIGURE 1. Procedural model for synthetic test development.

if a task has a significant affective component. If it does, the ultimate method of testing that task must be designed to accommodate that aspect of the task. In the case of performance under aversive conditions, this means that an ethically acceptable but psychologically effective simulation must be devised. Where the task requires an unalerted reaction to an unpredictable stimulus, or the voluntary selection of one of several competing response tendencies, the test method should be developed around an unobtrusive set.

Regardless of the affective nature of a task, the degree and type of skilled performance required must be evaluated. A task that is essentialy cognitive presents the least problem from the standpoint of efficient testing. Deciding, planning, evaluating, computing, and problem solving are examples of behaviors that normally can be group-tested using low fidelity stimulus simulation. The overt aspect of the response also is well suited to the group mode, in that it usually consists of a simple written or oral statement. If, however, a task involves perceptual or psychomotor elements, an acceptable substitute to performance testing is less straightforward.

Where perceptual and motor elements are judged to be of low skill -- that is, where the task can be performed by an untrained person with nothing more than procedural guidance -- one is led to the block in the model which states, "Test knowledge of sequence, procedure, and outcome using stimulus and response appropriate to verbal skills of examinees." The assumption is that ability to perform such tasks is almost entirely a function of <u>knowing</u> what to do, when to do it, and when it has been done correclty; actual execution of the steps in performance being within anyone's capability. Of course, to prevent test bias, stimulus

media and response modes must be selected with care. Verbal skill or, for that matter, any media-specific ability unrelated to task performance can spuriously effect test performance. If such effects can be neutralized in the design of test methods, knowledge testing offers an efficient means of evaluating performance on low-skill procedural tasks.

Job tasks entailing skilled perceptual or psychomotor elements must be treated differently. As indicated in the model, we must first determine if such a task can be meaningfully partitioned into subtasks or steps. If not -- if it is a highly integrated perceptual or motor skill -- then task integrity should be preserved, and testing should follow the tentative guidance given in one of the two remaining test method blocks in the diagram. The important requirement in testing a highly skilled perceptual task would seem to be high stimulus fidelity. In the case of skilled psychomotor performance, it is the fidelity of response-produced feedback stimuli that is critical if, as discussed above, negative transfer effects are to be avoided.

Where a task can be partitioned sensibly into subtasks or steps, the most difficult part of the task should be identified and a test method selected that is appropriate to that part. Three design options are possible at this point: one would be to limit testing of the task simply to a part-task test on the most difficult element; a second would be to test all task elements, but do so using the method necessary for the most difficult one; and, finally, a mixed-method approach could be taken in which the most difficult element were tested using one method, and remaining parts of the task tested

using another more efficient method. The option taken will depend on the task and its particular mix of elements. The last one mentioned, for instance, would seem well suited to lengthy procedural tasks in which only one step is skilled. Here the most efficient strategy might be to develop a part-task simulator to test performance on the skilled element, then test performance on the rest of the task in terms of knowledge of procedures.

Subtasks or steps that are time-shared are easily overlooked in analyzing task performance. These should be partialled out and viewed as skilled behavior for testing purposes. Since in most cases they will constitute the most difficult aspect of task performance, they should be so treated.

Evaluative Comment

In its present form the model is primitive. It does embrace the three most problematic and interactive aspects of the task-test domain: type of task, task element sampling, and test method. Moreover, the general framework seems useful as a way of codifying eventual guidelines for the development of synthetic tests. The rationale and definitions underlying each of the decision points need further development, and the method categories need expanding into hypothesized procedures. Yet conceptual work can go only so far without empirical data. Maturity of a synthetic approach to test development can be brought about best by an interplay of conceptual and experimental effort. If nothing else, in its current stage the model offers a number of interesting hypotheses.

Hypotheses relating to synthetic test approaches for psychomotor tasks seem especially important. The classification of tasks discussed in the first section of the report revealed that most 11D tasks involve things (equipment). And it is probably safe to speculate that many tasks in other combat arms MOS, if not in most Army jobs, are similarly oriented toward equipment. These kinds of tasks, moreover, are costly to test in the normal hands-on mode. Success of the Army's new Skill Qualification Testing Program for enlisted MOS will hinge largely on its ability to test efficiently these "hands'on" type tasks. Early experimental effort should therefore be devoted to test approaches implied in the model for both low and high skill psychomotor tasks. Pilot experiments presented in the following section were so designed, focussing on the paths leading to the second and fourth Method blocks in the model.

PILOT EXPERIMENTATION

Two parts of the preliminary model were selected for pilot exploration: 1) knowledge testing of low-skill psychomotor tasks, and 2) synthetic testing of skilled psychomotor tasks. Experiments carried out in these areas are described in this section of the report.

LOW-SKILLED PROCEDURAL TASKS

A manual task relatively unrestrained by time and composed of discrete steps, each of which can be performed with one or two guided practice trials, is considered to be a low-skilled psychomotor task. In other words, the more manipulative practice required for mastery, the more skilled the task. This is not to say that such tasks require no learning, since knowledge must be acquired of what steps to perform, with what result and in what order. The critical feature of this type of task, in fact, is that any skill involved appears to be predominantly mental. If so, task proficiency can be measured validly in a knowledge testing mode, given a test medium that is relatively neutral with respect to differences in mental ability.

Purpose

The purpose of this experiment was to assess four methods of knowledge testing in terms of their relative and absolute correlation with "hands-on" task proficiency for high versus low mental ability Ss.

The following research questions were of specific interest:

- Do the four types of knowledge test correlate with task mastery?
- Do the types of test differ with respect to how well they distinguish masters from nonmasters?
- Do the types of test distinguish masters from nonmasters equally well for high and low mental ability levels?
- Do the types of test tend to produce the same <u>kinds</u> of errors in predicting task mastery?

Method

<u>Test Development</u>. Tests were developed for three tasks: Installation of the Field Telephone (TEL), Setting up a Mechanical Ambush with the Claymore (AMB), and Disassembling the M-16 Rifle (RIF). The first two are clearly low-skilled tasks. Rifle disassembly, however, would be classified more accurately as moderately skilled, since some of the steps entail manipulations that are not easily mastered in one or two trials. Each task was analyzed into steps which were the referents for all test items. In addition to a performance (hands-on) test, four versions of a knowledge test were developed for each task. One version was a conventional multiple-choice test. The other three employed pictures in an effort to minimize literacy demands, but used different methods of eliciting task knowledge. A description of the four tests follows.

- <u>Written Choice (WC)</u>. This is a standard multiple-choice test consisting of one question for each step in the task.

A question focused on recognition of how a step is performed, when it is performed, or what its correct outcome is. Alternative answers to a question were limited to realistic options; unrealistic distractors were avoided. The test was scored by giving one point for each correct answer; seven was the maximum possible score for the TEL and AMB tasks, and eight the maximum for RIF.

- Picture Choice (PC). This method included the same questions as the Written Choice, but photographs were used in place of the printed word in presenting answer alternatives. The possible points and scoring procedure were the same as for WC.

- Picture Outcome (PO). In this method a photograph of the result of an improperly performed task was presented. Ss were instructed to inspect the picture and circle any errors. This type of test focuses on recognition of correct task outcome only. Test score was based on one point for each error circled, minus one point for each non-error circled. Total score was not allowed to go below zero. The possible range of scores was from 0 to 4 for TEL and RIF, and 0 to 3 for AMB.

- Picture Sort (PS). Photographs of steps in task performance, including both correctly and incorrectly executed steps, were used in this test method. The pictures were scrambled and presented to S with instructions to select the correct steps and place them in the order they should be performed. This method was considered to be the most comprehensive in its coverage of task knowledge; what steps to perform, and how and when to perform them are required knowledge. The method relies on recognition, as do the others, but all task elements are tapped and the guessing factor is minimized. Scoring was based on the award

> of one point for each picture or group of pictures representing a correct step performed in proper sequence. If two correct steps were in improper order, credit was withheld for the first step. Steps were judged to be improperly sequenced only if it were impossible or hazardous to perform them in that order. Maximum possible score was seven for TEL, and eight for AMB and RIF.

A complete set of tests, including the scoresheet for the performance test, on the telephone installation task is provided as an example in Appendix B.

It should be pointed out -- particularly to the reader who may be interested in adapting these test methods to other tasks -- that slight variation in test methods may be required by certain task characteristics. Three such characteristics surfaced during our test development efforts. The first pertained to <u>complexity</u> of steps in task performance, where complexity refers to the number of sub-steps, or closely associated manipulations, required to perform part of the task. For example, in clearing the rifle seven substeps are performed -- pull charging handle, press bolt latch, engage bolt latch, return charging handle to forward position, inspect receiver, place selector lever in SAFE, depress bolt latch. That seven-part step is harder to capture pictorally than a simple step like setting the circuit selector switch on the field telephone. Full performance of a simple step can be represented easily in one photograph; full performance of a complex step usually requires several photographs to represent the actions. In an effort to keep the alternatives as simple as possible

on the PC test for rifle disassembly, the first item addressed only the last part of the step. Even then the alternatives required more than one photograph (Appendix C). That item is the only exception to the rule tentatively applied in preparing the PC tests -- represent a step with one photograph.

The second characteristic that precluded uniform application of test development rules pertained to the <u>spatial</u> <u>relationship</u> among equipment components. An important decision when preparing the PO tests was to select a camera view that showed enough detail of each component and enough of the relationship among the components for examinees to evaluate the task product. For most "thing-oriented" tasks, this decision is uncomplicated. Usually the equipment is compact enough that a photograph of the equipment will show the location and detail of each relevant component. However, a detailed PO for the mechanical ambush could not be captured in a single photograph since task product consisted of widely separated components. The first attempt to show the product consisted of a schematic of the layout with close-up photographs of each component. Since preliminary tryouts of the test indicated that examinees still had difficulty maintaining perspective of the layout, the picture was revised. A wide-view photograph of the layout was used for the background, and blow-up photographs of individual components were attached (Appendix D).

A third characteristic was revealed during development of the WC tests. All knowledge -- especially nomenclature -- considered irrelevant to task performance was judiciously excluded from test

content. Yet in a written mode it is difficult to ask questions about task procedures without using nomenclature. Equipment components for the TEL and AMB tasks were referred to in terms familiar to most soldiers, but rifle parts were not. This problem was resolved by giving each examinee a labelled picture of the rifle for reference.

These variations in test method are not considered preemptive to the design of this research. They do, however, represent the kinds of issues that will temper ultimate procedures for synthetic test development.

Subjects. Thirty-seven soldiers from units at Fort Knox were tested. They were chiefly from combat arms MOSs and ranged in grade from E-2 to E-6. For the purpose of study design, Ss were in two mental ability (MA) groups: GT over 110 (high MA), and GT under 90 (low MA)[1]. Twenty Ss were in the high MA group and 17 were in the low.

Procedure. On arrival at the test site, the project was explained briefly to Ss. What was said to them took the following general form:

> We are working on a project to evaluate several different methods of testing. You will take a hands-on test for three tasks. Then you will take four other kinds of tests for each task. After the test we will ask your opinion of it. This is not an MOS test, so there is no reason for you to be nervous. But the project is very important so, of course, we expect you to do as well as you can on every test.

---

[1]The GT (General-Technical) is a combination of scores on a verbal and a quantitative aptitude test. It is considered to be the best indicator of general mental ability in the Army Classification Test Battery.

All testing was done individually and began with administration of the hands-on test. At this point some Ss received training on the task before going on to the knowledge tests. This was done to control the range of task mastery within the two MA groups. The intention was to create a rectangular distribution of mastery, with approximately a third of each MA group being wholly unqualified on a task, a third being partially qualified, and a third full masters. This approach worked well at the full mastery level since only one S could perform a task (TEL) without further training. Thus, 7 masters were created in each MA group by training them to pass the three hands-on tests. The approach did not work as well within the nonmastery range since most Ss could perform a few steps in the TEL and RIF tasks; only with the AMB task were any Ss trained to partial mastery.

Once an S had completed the hands-on test for a task, he was given the four knowledge tests successively. The order of test administration was counterbalanced over Ss and, with exception of the three additional high-MA Ss, the orders were the same for both MA groups.

In addition to test performance, Ss were asked their opinions of the methods. Two questions were asked after each test:

- "Do you think this test is a good way to find out if a soldier can (task statement)?" A five-point response scale was used, ranging from, "not good at all" to "very good."
- "Do you think your performance on this test gives a fair picture of your ability to (task statement)?" This question was answered "yes" or "no."

After an S had taken all tests for a given task, he was asked to rank them from 1 to 5 with respect to the first question above. Also, after testing was completed on all tasks, he was asked to rank the five test methods overall.

Scores on the 15 tests -- one hands-on and four knowledge tests for each of three tasks -- and Ss ratings comprised the data to be analyzed.

Results

Continuous score correlations between knowledge test and hands-on performance for the three tasks are shown in Table 2 for the two levels of mental ability and for the total sample. With few exceptions the correlations are both statistically and practically significant. They are uniformly higher, regardless of test method, for the TEL and AMB tasks than for RIF, indicating that rifle disassembly is somehow different from the other tasks; a difference attributable perhaps to a more skilled motor component.

Comparison by type of knowledge test, for the total sample and for total performance on the three tasks, indicates that the Written Choice, Picture Choice and Picture Outcome correlate equally well (.83, .80, and .84 respectively) with hands-on performance. The Picture Sort method yields a somewhat smaller overall relationship (.58), although the reduction is attributable to the near-zero correlation for the RIF task. The trend toward higher correlations for total score than for task scores relects a tendency for inter-

TABLE 2

CORRELATIONS BETWEEN PERFORMANCE AND KNOWLEDGE TEST METHOD FOR HIGH AND LOW MENTAL ABILITY GROUPS

MENTAL ABILITY GROUP	N		KNOWLEDGE TEST METHOD: WRITTEN CHOICE				PICTURE CHOICE				PICTURE OUTCOME				PICTURE SORT			
			TASK [a]: TEL	AMB	RIF	TOT	TEL	AMB	RIF	TOT	TEL	AMB	RIF	TOT	TEL	AMB	RIF	TOT
HIGH	20	r	.69	.55	.17	.79	.71	.66	.47	.82	.76	.77	.31	.78	.70	.62	-.31	.52
		$\bar{X}$	4.90	5.20	5.40	15.45	5.05	5.40	6.40	16.85	2.40	1.60	3.75	7.75	4.85	5.80	5.75	16.40
		s	1.52	1.74	1.67	3.60	1.76	1.57	.94	3.53	1.23	1.14	.44	2.34	1.81	2.33	1.52	4.43
LOW	17	r	.73	.82	.75	.90	.80	.76	.51	.80	.68	.74	.65	.90	.79	.56	.29	.69
		$\bar{X}$	4.76	4.82	4.41	14.06	5.06	5.00	5.82	15.94	2.35	1.35	3.18	7.12	4.59	4.41	5.29	14.29
		s	1.48	1.67	1.70	3.90	1.64	1.54	1.67	3.72	1.11	1.11	1.33	2.91	1.54	2.18	1.61	4.19
TOTAL	37	r	.71	.67	.49	.83	.75	.70	.51	.80	.72	.74	.55	.84	.72	.55	.04	.58
		$\bar{X}$	4.84	5.03	4.94	14.78	5.05	5.22	6.13	16.43	2.38	1.49	3.48	7.46	4.73	5.16	5.54	15.43
		s	1.48	1.69	1.73	3.71	1.68	1.55	1.34	3.59	1.16	1.12	.99	2.60	1.68	2.34	1.56	4.39

[a] *TEL = Installing Field Telephone*

AMB = Installing Mechanical Ambush with Claymore Mine

RIF = Disassembling M16 Rifle

TOT = Total Performance on the Three Tasks

correlations among tasks to be lower for a knowledge test than for the hands-on criterion.[2]

Further analyses of the effectiveness of the different knowledge tests to distinguish masters from nonmasters, both within and between levels of mental ability, were carried out by analysis-of-variance. This is a reasonable way to examine the data, since mastery level was more of a manipulated "treatment" effect than a natural variate. Knowledge test performance, summed over tasks, of masters and nonmasters by mental ability level is shown in Table 3. All test methods did not have the same scale of measurement, so an ANOV (Winer, 1962) was performed on each method. Results of the four unweighted means ANOV are summarized in Table 4 and shown graphically in Figure 2. A clear and substantial main effect is revealed for mastery level, which merely represents the high correlations between knowledge test and task performance already mentioned. The size of this main effect for Picture Outcome relative to other test methods is worthy of note. The graphs in Figure 2 indicate that masters tend to average about five points higher than nonmasters on all tests, even though the <u>potential</u> range of performance on PO is only half that of the other tests. This would imply that a longer test would produce greater improvement in discrimination between masters and nonmasters for PO than for the other methods.

---

[2] The reader will recall that, by design, the same people were masters on all tasks (had maximum criterion scores) although nonmasters varied in degree of nonmastery from task to task.

TABLE 3

KNOWLEDGE TEST PERFORMANCE (MEANS AND STANDARD DEVIATIONS) OF MASTERS AND NONMASTERS BY TEST METHOD AND MENTAL ABILITY LEVEL

MASTERY LEVEL	MENTAL ABILITY		TEST METHOD WRITTEN CHOICE	PICTURE CHOICE	PICTURE OUTCOME	PICTURE SORT
MASTERS	HIGH	$\bar{X}$	18.71	20.28	10.29	18.71
		s	2.98	1.60	.76	3.25
		N	7	7	7	7
	LOW	$\bar{X}$	17.86	19.29	10.14	17.57
		s	1.95	2.10	1.21	3.41
		N	7	7	7	7
NONMASTERS	HIGH	$\bar{X}$	13.69	15.00	6.38	15.15
		s	2.56	2.80	1.61	4.56
		N	13	13	13	13
	LOW	$\bar{X}$	11.30	13.60	5.00	12.00
		s	1.83	2.59	1.41	3.06
		N	10	10	10	10

TABLE 4

ANOV SUMMARIES OF THE EFFECTS OF TASK MASTERY (*M*) AND MENTAL ABILITY (*A*) ON KNOWLEDGE TEST PERFORMANCE

TEST METHOD	SOURCE	SS	df	MS	F
WRITTEN CHOICE	M	289.853	1	289.853	46.10**
	A	22.691	1	22.691	3.61
	$M \times A$	5.126	1	5.126	.82
	Error	207.466	33	6.287	
PICTURE CHOICE	M	260.120	1	260.120	43.73**
	A	12.357	1	12.357	2.08
	$M \times A$	.364	1	.364	.06
	Error	196.273	33	5.948	
PICTURE OUTCOME	M	177.034	1	177.034	95.38**
	A	5.060	1	5.060	2.73
	$M \times A$	3.271	1	3.271	1.76
	Error	61.2483	33	1.856	
PICTURE SORT	M	180.178	1	180.178	12.73**
	A	39.781	1	39.781	2.81
	$M \times A$	8.733	1	8.733	.62
	Error	466.939	33	14.150	

** $p < .01$

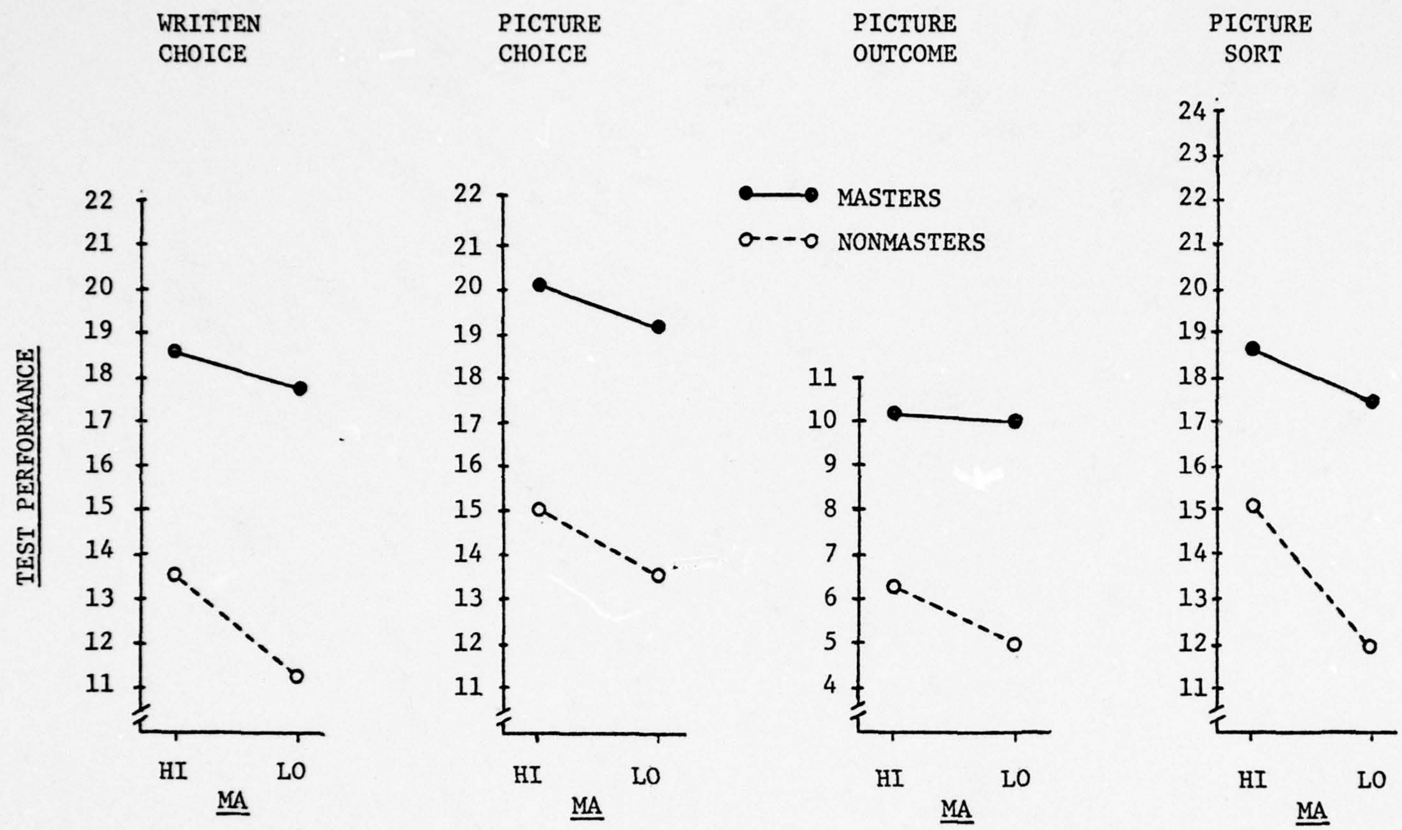


FIGURE 2. Mean performance of masters and nonmasters by mental ability (MA) level for the four knowledge test methods.

Performance on the knowledge tests tended to be lower for low mental ability Ss than for high, as indicated by the slope of the curves in Figure 2. The difference is small, and in fact not statistically reliable according to the separate ANOVs. However, when performance was converted to standard scores within test method and aggregated over methods, the mental ability factor is marginally significant ($p > .05$). Moreover, the difference appears to be relatively constant over test methods (Figure 3), suggesting that no one method is superior in neutralizing mental ability differences.

One of the more interesting features of the data (Figure 2) is the trend, however slight, toward a larger difference between masters and nonmasters in the low MA group. This indicates a slightly higher correlation between knowledge test and task performance for low mental ability Ss, a tendency also observed in Table 2 where for 9 of the 12 method/task combinations the correlation with mastery was higher within the low mental ability group. Note that this is not a statistically reliable phenomenon, but it suggests an interesting hypothesis: knowledge based tests predict task performance better among people of moderate to low mental ability than among those of high mental ability.

Validity in a strict correlational sense does not tell the whole story, however. The type of prediction or classification error is of practical interest. By converting knowledge test performance to pass-fail scores and arraying them against the master-nonmaster criterion, four-fold tables were generated from which the incidence of false negative (masters who failed the test) and false-positive (nonmasters

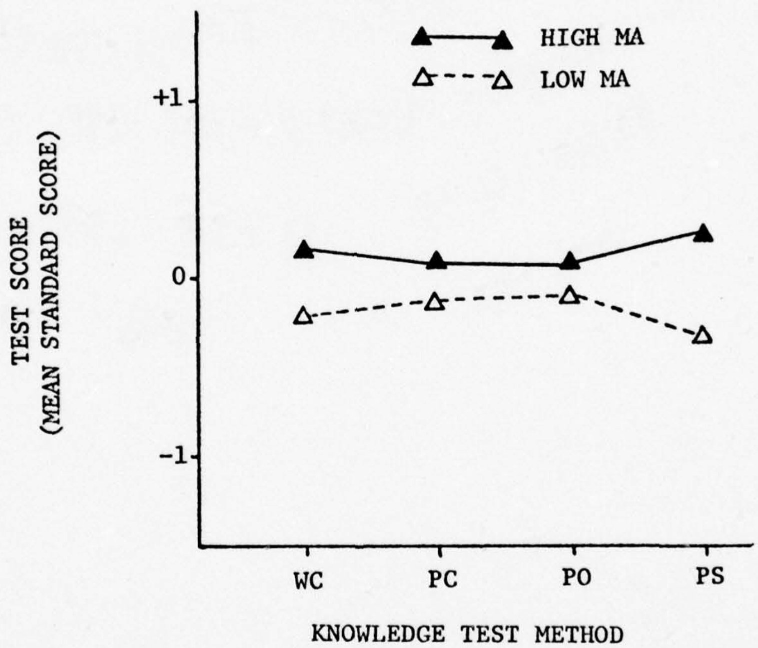


FIGURE 3. Mean standard score performance of high and low mental ability (MA) groups for the four knowledge test methods.

who passed the test) classification errors were determined. The correlation and amount of classification error, of course, depend on the standard used in scoring pass-fail. Classification error was tabulated for a standard of full mastery on the knowledge test (pass = all items right) and again for a standard of part mastery (pass = no more than one item wrong). The results are shown in Table 5 for high and low mental ability groups and for the total sample. With exception of the Picture Outcome method, classification error is somewhat less using the more liberal part mastery criterion on the knowledge tests. Total error tended to run about 25% on the average, reaching a low of 16% for the Picture Choice method with the criterion of part mastery. Of particular interest is the distribution of total error between false-positive and false-negative categories. As the standard for passing a predictor measure is relaxed, the number of false-positives generally increases. The optimal ratio of the two types of error is a moot point, and will depend largely on how test scores are to be used. But if test fairness is the goal, then minimizing the number of false-negatives should be the objective. The relative number of false-negatives, moreover, should be the same for groups differing in mental ability (or any other ability correlated with test score but unrelated to criterion performance). Comparing high and low MA groups we find a small but consistant tendency toward more false-positives among the high MA's, and more false-negatives among the low. This trend was evaluated by *Chi-square* analysis of the difference in type of

TABLE 5

AVERAGE[a] PERCENT CLASSIFICATION ERROR AS A FUNCTION OF KNOWLEDGE TEST METHOD AND LEVEL OF MENTAL ABILITY

TEST STANDARD	MENTAL ABILITY GROUP	WC FN	WC FP	WC TOT	PC FN	PC FP	PC TOT	PO FN	PO FP	PO TOT	PS FN	PS FP	PS TOT
FULL MASTERY	HIGH	18	05	23	13	07	20	05	15	20	23	03	26
	LOW	27	02	29	25	04	29	16	08	24	33	00	33
	TOTAL	22	04	26	19	05	24	10	12	22	28	02	30
PART MASTERY	HIGH	07	13	20	02	13	15	00	32	32	15	17	32
	LOW	18	02	20	08	10	18	04	20	24	22	02	24
	TOTAL	12	08	20	04	12	16	02	26	28	18	10	28

(Column heads: KNOWLEDGE TEST METHOD — WC, PC, PO, PS; CLASSIFICATION ERROR[b] — FN, FP, TOT)

[a] *Averaged over the three tasks.*

[b] *FN = False Negatives (masters who failed knowledge test)*
FP = False Positives (non-masters who passed knowledge test)
TOT = Total Classification Error

TABLE 6

CHI SQUARE OF THE DIFFERENCE IN TYPE OF CLASSIFICATION ERROR BETWEEN HIGH AND LOW MENTAL ABILITY GROUPS BY TEST STANDARD AND TEST METHOD

TEST STANDARD	KNOWLEDGE TEST METHOD WC	PC	PO	PS
FULL MASTERY	1.33	1.54	4.19*	2.26
PART MASTERY	7.22*	2.49	3.38*	6.30*

* $p < .10$

classification error between high and low MA groups, and is shown in Table 6 by test method for each standard of test "mastery." Observed *Chi-squares* were tested at the 10% level of significance, which provides for a conservative decision with respect to accepting the null hypothesis of no difference between groups in distribution of classification error. Type of classification error produced by the knowledge tests does appear to interact with mental ability. Although the number of cases underlying the analysis are too few to warrant firm conclusion, indications are that if one were interested in minimizing the incidence of false-negatives (i.e., the part mastery standard), the Picture Choice method produces the most equitable results for both mental ability groups.

Personal Preferences for Test Methods. Ss' opinions of the test methods were solicited after each test was administered and again when all testing was concluded. Responses at the two points in time were similar, so only the final ratings are reported here. Ss were asked to rank the five methods (including the hands-on criterion test) from highest to lowest in terms of the question, "Do you think this test is a good way to find out if a soldier can ... (e.g., set up a mechanical ambush with a Claymore?") Rankings were done separately for each task. Overall mean preference was highest for the hands-on method of performance testing, as might be expected (Tables 7 & 8). Differences in preference for the four methods of knowledge testing were less pronounced, although the Picture Choice consistently received higher average ranking regardless of the referent task or rating

TABLE 7

MEAN ORDER OF PREFERENCE [a] BY TASK FOR
THE HANDS-ON AND KNOWLEDGE TEST METHODS

TASK	HANDS-ON	TEST METHOD WC	PC	PO	PS
TEL	1.14	3.92	3.03	3.58	3.33
AMB	1.25	3.78	2.94	3.72	3.31
RIF	1.08	3.67	3.14	3.47	3.64

[a] *The lower the number the higher the preference.*

TABLE 8

MEAN ORDER OF PREFERENCE BY SUBGROUP
FOR THE HANDS-ON AND KNOWLEDGE TEST METHODS

SUBGROUP	HANDS-ON	TEST METHOD WC	PC	PO	PS
MASTERS	1.00	3.85	3.08	3.38	3.69
NON-MASTERS	1.30	3.83	2.65	3.87	3.35
HIGH MA	1.35	4.10	2.65	3.70	3.20
LOW MA	1.06	3.50	3.00	3.69	3.81
TOTAL	1.19	3.83	2.81	3.69	3.47

subgroup. Overall, the hands-on method was first, Picture Choice second, Picture Sort third, Picture Outcome fourth, and Written Choice last in average order of preference.

Discussion

A number of interesting though tentative findings emerged from this study. The small sample of people and tasks certainly limits generality of the results, and the following interpretation and conclusions should be so tempered.

The data strongly support the hypothesis that performance on low-skill procedural tasks is mediated by knowledge. Correlations between task knowledge and task performance were high, particularly for the two procedural tasks with the lowest skill requirements. The correlations reached as high as .75 in spite of the fact that the range of possible test performance seldom exceeded seven points. When performance was aggregated over tasks, the correlations tended to be more on the order of .80.

Substantial differences among methods of knowledge testing were not found. The conventional written multiple-choice test did essentially as well as the pictorially based methods in distinguishing masters from nonmasters. (In this connection, however, it should be noted that test questions were carefully directed at steps necessary in task performance, and did not include those marginally relevant knowledge items often found on such tests.) Failure of the Picture Sort tests to correlate higher with performance was an unexpected

result. This method was designed to tap more fully all knowledge aspects of task performance, including recognition of the steps, their correct outcome, and sequence. In so doing, however, it may well have become the most demanding test technique from the standpoint of method-specific mediation requirements; that is, the examinee must first analyze what he does in performing the task, and then synthesize it a step at a time by sorting through a large number of pictures more or less representative of his mental images of the task. That kind of abstract manipulation probably taxes the intellectual and visualization abilities more than we originally anticipated. In support of this speculation, there was some indication that Ss in the low mental ability group had more trouble with this test method than with others (Figure 2). The written and pictorial multiple-choice tests, though more dependent on literacy, represent a culturally familiar method. The Picture Outcome method appears to be the simplest in the sense of minimizing both literacy and method-specific mediational demands, and is certainly worthy of further study and development as an efficient method of knowledge testing.

Correlations between knowledge and performance were not significantly different for high versus low mental ability Ss. Yet there was a slight but noticeable trend toward larger correlations within the low mental ability group. The possibility that knowledge measures -- including the standard multiple-choice test -- are better predictors of task mastery for those of below average mental ability

is intriguing. If true, we need to reevaluate the popular notion that knowledge tests of manual performance are unfair to those less apt in the academic skills of reading, writing and symbol manipulation. The notion _is_ probably valid, but it may be so for reasons quite different than normally offered. Knowledge tests apparently are good predictors of performance on low-skill procedural tasks among people of low to moderate mental ability. The unfairness lies not in the inability of this group to use a knowledge testing medium, but in the tendency of brighter people to _over_ use it. The hypothesis here is that some minimum level of ability, whether innate or acquired, is necessary to handle the symbolic and semantic demands of a knowledge test; but beyond that level, correlated factors such as test-wiseness begin to moderate the true relationship between task knowledge and performance. Two additional features of the data tend to support this speculation: a) higher average knowledge test scores for the high mental ability group, and b) relatively more false-positive errors in predicting mastery among this group.

If one were urged to recommend, on the basis of this study, a method of testing knowledge on low-skill procedural tasks, the Picture Choice would probably have to be named. The data are certainly not conclusive, but this method came the closest to meeting the overall validity criteria: it demonstrated a high correlation with hands-on task performance; the correlation was relatively constant over the range of mental ability; and, the distributions of classification error were more nearly proportional for the two levels of mental ability. Moreover, the Picture Choice method was second only to the hands-on test in examinee preference.

SKILLED MOTOR TASKS

Tasks involving skilled manual responses present a special challenge to synthetic testing. Short of full task enactment in a realistic setting and with actual equipment, what options are available? Knowledge tests can safely be ruled out since acquisition of fine-grained motor responses does not take place at the cognitive level. At intermediate or low levels of skill acquisition task knowledge may correlate with performance, but it is most likely a poor gauge at the mastery or near-mastery level.

Potential difficulties associated with task simulations in testing have been discussed. It seems reasonable to expect negative transfer from task performance to test performance when masters or near-masters are tested with anything short of high fidelity simulations. Because of this, good quality simulation of response media -- especially of response-produced feedback stimuli -- is recommended for synthetic tests of skilled motor behavior. The problem with this recommendation is that little gain in testing efficiency can be expected if the fidelity requirement is extreme. In weapon firing, for example, if the interaction between the "feel" of the weapon, the discharge of the round, and target reaction must be simulated with near maximum fidelity in order to achieve test validity, then it is unlikely that much is to be gained in economy over testing on a live-fire range. Some feedback characteristics would have to be substantially degraded before much testing economy would be realized. On the other hand, if certain characteristics can be degraded without

effectively changing the task (and thereby inducing negative transfer), testing could be done less expensively. What aspects of the feedback stimulus complex can be inexpensively simulated in valid tests of motor skills is an experimental question.

Another approach to synthetic testing of motor skills, not reflected in the preliminary model, is worthy of exploration. It is based on two assumptions: a) in any motor skill there is at least one key response feature that distinguishes masters from non-masters; and, b) in eliciting even a facsimile of that feature through part-task simulation, the response of a master will differ generically from that of a nonmaster. The important implication of this hypothesis, if true, is that low level simulations may be effectively used as test media, since the elicited response need not be the same as that in actual task performance; it is only necessary that the test response reliably characterize task masters as opposed to nonmasters.

Three small demonstration studies were designed to explore these assumptions. Apparatus problems prevented completion of one of them, but its methodology seemed worthy of description.

Study I

The first study attempted to demonstrate the effectiveness of a synthetic test method in which aspects of the feedback complex were degraded. The task was aim-firing the 45 caliber pistol. With exception of the round explosion which was not simulated, intra-

task feedback (weapon "feel") was maintained at maximum fidelity by using an actual weapon. External feedback, target reaction, was simulated at a low level. Ambient stimuli were totally excluded from the simulation.

The method involved "firing" the pistol at a scaled-down paper target approximately two inches from the end of the barrel.[1] The target, a 3" x 5" card, with a one millimeter dot as an aim point, was mounted on the wall at shoulder level. An ordinary sharp pencil, taped to fit the bore and inserted eraser-first into the barrel, served as the "round." When fired from a distance of about 1" the pencil was driven by the firing pin to the target, leaving a mark on impact.

Six Ss, all members of the research staff, participated in the experiment. Three were ex-Army officers, each of whom had several years' experience firing the .45 pistol. The other 3 Ss were considered inexperienced; two had no experience with the weapon, and the other had once fired it for familiarization. Each S completed six trials on the synthetic test. A trial consisted of a group of three "rounds" fired at one aim point. The three-shot cluster was scored by measuring and summing distances between the hit points. These totals were averaged over the six trials and are shown in Table 9. Means for the experienced shooters were uniformly lower than for inexperienced, the group means being respectively 6.2 mm and 10.5mm.

---

[1]This method is used occasionally in the military as a field-expedient training technique.

TABLE 9

SCORES[a] FOR EXPERIENCED AND INEXPERIENCED SHOOTERS ON A SYNTHETIC TEST OF PISTOL MARKSMANSHIP

EXPERIENCED					INEXPERIENCED			
S	1	2	3	TOTAL	4	5	6	TOTAL
$\overline{X}$	4.96	6.67	7.01	6.21	9.42	14.0	8.18	10.53
s	4.04	2.94	1.67	1.09	4.05	6.62	3.92	3.06

[a]Distance in mm between rounds in a three-shot group average over six trials

Although scores differed in the expected direction, the differences were not evaluated statistically, since the N was small and the study rather informal. Moreover, to validate the test method properly, it would be necessary to replicate the study on true masters (men who had recently fired "Expert," "Sharpshooter" or "Marksman" on their annual qualification) versus experienced nonmasters (those who scored "Unqualified"). In spite of these shortcomings the results are encouraging. It may well be possible to test this type of tracking skill validly with substantial reduction in external feedback fidelity.

Study II

The intention in this study was to determine whether a pattern of response or "signature" distinguishing masters from nonmasters could be identified for a motor skill. The referrent task again was pistol firing, and the response feature of interest was barrel

movement in aiming. A skilled shooter's hand may not be steadier than that of the less skilled. Possibly the two differ only in how they cope with the unsteadiness. Rather than fighting involuntary hand movement, the skilled marksman may direct it in a controlled pattern and time trigger squeeze accordingly. In either case, whether less movement or a patterned movement, some type of signature should emerge, and it should be detectable at a low level of simulation.

To explore this possibility a simple simulator was constructed. The apparatus consisted of a dowel rod cantilevered through a heavy rubber diaphram which was mounted over a hole in a board. The barrel of a mock or real pistol could be attached to the front end of rod, and a marking pen to the other. The pen tracked barrel motion on a moving strip of paper as one sighted on a target just above the hole. The apparatus, unfortunately, did not work out. Barrel movement was not sufficiently amplified in the recording to be useable. Hopefully, some other simple approach can be found that will not involve sophisticated mechanical or electronic hardware, since potential efficiency of the method seems so important.

Study III

The possibility of a characteristic part-task response that distinguishes masters from nonmasters can be carried even further down the simulation scale. Although the notion of testing knowledge <u>about</u> skilled performance has been rejected, it is possible that

cognitively mediated aspects of a motor skill can be identified, measured, and validated against performance. Sizeable gains in testing efficiency could be realized if this were the case.

A brief test of this possibility was carried out using the task of typing. Three "master" typists (greater than 50 words per minute) and 3 nonmasters (less than 30 words per minute) were tested on their ability to call out the hand sequence (e.g., "left, right, right, left,...") used in typing a given word. Fourteen 5 letter words of 1 or 2 syllables were individually printed on flash cards. Each S was given five practice trials to become familiar with the task, before responding to the 14 item test. Time and errors were recorded for each trial.

Masters averaged an error on 2 of the 14 words, and nonmasters on 4.3 of 14. Overall, errors were too few to warrant analysis. Mean time scores are shown in Table 10. The group means differed in the

TABLE 10

MEAN TIME SCORES (SEC.) FOR MASTERS AND NONMASTERS IN REPORTING HAND SEQUENCE FOR TYPING WORDS

MASTERS					NONMASTERS			
S	1	2	3	TOTAL	4	5	6	TOTAL
$\bar{X}$	4.6	3.7	3.2	3.9	10.1	7.4	4.3	7.2
s				1.2				4.0

expected direction, but were not significant. Though statistical significance could probably be attained with a larger sample, practical significance might remain questionable since one of the nonmasters did better on the test than one of the masters; had a mastery standard been established for the synthetic test, at least one classification error would have resulted. It is interesting to speculate here about the possibility of negative transfer inhibiting performance of master typists on such a test. In the advanced stages of typing skill development, individual letters lose their distinction and the typist begins to type whole words. Once this higher level of skill is automatized, it may be difficult for a typist to disintegrate quickly the hand or finger sequence used in typing individual letters of a word. This is a good example, perhaps, of how skilled task performance can actually interfere with performance on a synthetic test.

Though short on substance, these final studies point to a feature of tests not formerly included in the definition of efficiency: fairness of test medium. A synthetic test must be feasible and valid; high correlation with performance is an imperative. But test developers must also guard against the bias in errors of prediction that interactive effects of ability and media can produce. Paper and pencil, symbolic displays, verbalization, apparatus -- any medium used synthetically -- can inflate or inhibit the performance from which one wishes to generalize. Fairness is achieved through media that do not favor abilities unnecessary for task performance. Identifying such neutral media may be the most subtle and challenging mission of synthetic testing research.

REFERENCES

Ammerman, H.L. & Melching, W.H. The Derivation, Analysis and Classification of Instructional Objectives, HumRRO Technical Report 66-4, May 1966.

Baker, R.A., Cook, J.G., Warnick, W.L., & Robinson, J.P. Development and Evaluation of Systems for the Conduct of Tactical Training at the Tank Platoon Level, HumRRO Technical Report 88, April 1964.

Baldwin, R.D. Capabilities of Ground Observers to Locate, Recognize, and Estimate Distance of Low-Flying Aircraft, HumRRO Technical Report 73-8, March 1973.

Cox, J.A., Wood, R.O., Jr., Boren, L.M., & Thorne, H.W. Functional and Appearance Fidelity of Training Devices for Fixed-procedure Tasks, HumRRO Technical Report 65-4, June 1965.

Daugherty, D.J., Houston, R.C., & Nicklas, D.R. Transfer of Training in Flight Procedures from Selected Ground Training Devices to the Aircraft, Technical Report NAVTRADEVCEN 71-16-16, U.S. Naval Training Device Center, Port Washington, New York, September 1957.

Frederiksen, N., et al. "The In-Basket Test," Psychological Monographs: General and Applied, 1957, 71(9, Whole No. 438).

Glazer, R., et al. "The Tab Item," Educ. and Psych. Meas., 1954, 14, 283.293.

Mackie, R.R. & Harabedian, A.A. A Study of Simulation Requirements for Sonar Operator Training, Technical Report NAVTRADEVCEN 1320-1, Human Factors Research, Inc., Los Angeles, California, March 1964.

Osborn, W.C. "An Approach to the Development of Synthetic Performance Tests for Use in Training Evaluation," paper for 12th Annual Military Testing Association Conference, French Lick, Indiana, HumRRO Professional Paper 30-70, December 1970.

Osborn, W.C., Harris, J.H., & Ford, J.P. "Functionally Integrated Performance Testing," paper for the 16th Annual Military Testing Association Conference, Oklahoma City, Oklahoma, October 1974.

Osgood, C.E. Method and Theory in Experimental Psychology, London, Oxford University Press, 1953.

Prophet, W.W. & Boyd, H.A. Device-task Fidelity and Transfer of Training: Aircraft Cockpit Procedures Training, HumRRO Technical Report 70-10, July 1970.

U.S. Department of Labor. Dictionary of Occupational Titles, 3rd edition, Vol. II, Government Printing Office, Washington, D.C., 1965.

U.S. Department of Army. "Training Manager's Handbook for Situational Training," Draft Manual, U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington, Virginia, 1973.

Weitz, J. and Adler, S. "The Optimal Use of Simulation," Journal of Applied Psychology, 58, 1973, 219-224.

Winer, B.J. Statistical Principles in Experimental Design, New York, McGraw-Hill, 1962.

APPENDIX A

TESTING PROBLEMS IDENTIFIED FOR A SAMPLE OF 100 RECONNAISSANCE SPECIALIST JOB TASKS

TESTING PROBLEMS IDENTIFIED FOR A SAMPLE
OF 100 RECONNAISSANCE SPECIALIST JOB TASKS

	TASK CONDITIONS							TASK BEHAVIOR		
	SCARCE			DANGEROUS	VARIABLE		LATENT	LONG PROCESS	TRANSIENT PROCESS	AFFECTIVE
TASK	Equipment	Terrain	Personnel		Surround	Display				
1. Navigate from one point on the ground to another using a strip map		X			X	X		X	X	
2. Give CBR alarm				X	X		X			
3. Put on protective mask				X						
4. Maintain the gas particulate unit of a command and reconnaissance carrier (M114A1)	X					X				X
5. Evaluate terrain using an aerial photo as a supplement to a topographic map/pictomap										
6. Determine the scale of an aerial photograph										
7. Plan use of available training time										
8. Evaluate training test results										
9. Assist/prepare facilities for inspection	X					X		X		
10. Perform the duties of a member of an interior guard	X			X			X	X	X	
11. Plan escape as a PW	X	X						X		
12. Identify enemy vehicles and equipment	X	X								
13. Process captured documents and materials										
14. Maintain light discipline to reduce the danger of detection.							X			
15. Maintain fire discipline to reduce the danger of detection.							X			
16. Install field telephones	X	X								
17. Communicate information over tactical wire and FM radio nets	X								X	
18. Leave radio communications net.	X								X	
19. Apply low-level anti-jamming procedure	X								X	X
20. Install hot loop wire communication	X									
21. Prepare written messages										
22. Camouflage weapons	X					X				X

Continued

TESTING PROBLEMS IDENTIFIED FOR A SAMPLE
OF 100 RECONNAISSANCE SPECIALIST JOB TASKS

TASK	TASK CONDITIONS							TASK BEHAVIOR		
	SCARCE Equipment	Terrain	Personnel	DANGEROUS	VARIABLE Surround	Display	LATENT	LONG PROCESS	TRANSIENT PROCESS	AFFECTIVE
23. Enforce camouflage discipline..	X						X			
24. Construct bunkers..............	X							X		
25. Perform preventive maintenance on fuel stoves & lanterns......	X					X				X
26. Perform operator's checks & services on generator sets.....	X					X				X
27. Service an M16A1 rifle magazine						X				X
28. Conduct M16A1 rifle marksmanship training.................	X	X	X					X	X	
29. Engage a target during periods of limited visibility with an M16A1.........................		X		X	X				X	
30. Apply immediate action to reduce a stoppage in an M203 grenade launcher..............	X	X		X			X		X	
31. Disassemble/assemble an M79 grenade launcher..............	X								X	
32. Adjust fire of an M79 grenade launcher......................	X	X		X						
33. Correct malfunctions in an M60 machinegun....................	X					X				
34. Service the M2HBTT machinegun..	X					X				X
35. Engage targets with the M2HBTT machinegun mounted in an M114..	X	X		X					X	
36. Assist men in handling natural fears in avoiding panic........			X	X		X	X		X	
37. Receive and orient newly assigned unit personnel........			X			X				
38. Prepare enlisted efficiency reports......................										
39. Investigate complaints.........			X			X		X		
40. Counter disruptive influences and acts......................			X	X		X			X	
41. Encourage others to enlist in the Army......................			X			X			X	X
42. Respond to ground guide signals while driving a tracked vehicle	X	X							X	
43. Engage a target with an M60 machinegun using an AN/PVS-2...	X	X		X					X	

Continued

TESTING PROBLEMS IDENTIFIED FOR A SAMPLE
OF 100 RECONNAISSANCE SPECIALIST JOB TASKS

TASK	TASK CONDITIONS							TASK BEHAVIOR		
	SCARCE Equipment	Terrain	Personnel	DANGEROUS	VARIABLE Surround	Display	LATENT	LONG PROCESS	TRANSIENT PROCESS	AFFECTIVE
44. Mount/dismount AN/PVS-2 scope to caliber .50 machinegun......	X									
45. Store cleaned AN/PAS-6.........	X								X	
46. Store/carry AN/PVS-2...........	X								X	
47. Select explosives appropriate to mission....................										
48. Prepare a shaped charge for detonation.....................	X			X						
49. Destroy a mine in place........	X	X		X						
50. Disassemble mine detector ANPRS-7.......................	X								X	
51. Construct a double apron fence.	X	X	X			X		X		
52. Breech a mine field...........	X	X		X				X	X	
53. Lead a security patrol.........		X	X	X				X	X	
54. Determine the enemy's strength and disposition...............		X	X							
55. Supervise a route reconnaissance	X	X	X	X				X	X	
56. Restore M72A2 law to carrying configuration.................	X	X		X					X	
57. Adjust indirect fire using creeping method...............	X	X	X	X						
58. Traverse terrain dismounted in a tactical situation..........		X	X	X				X	X	
59. Coordinate unit defense plans with adjacent units...........			X							
60. Take passive measures to prevent detection by enemy aircraft.	X	X				X				X
61. Clear fields of fire...........	X	X				X				
62. Perform quartering party functions.....................	X	X	X					X	X	
63. Conduct administrative movement	X	X	X					X	X	
64. Recommend allocation of ammunition and equipment.............										
65. Inspect training..............	X		X			X		X		
66. Recommend personnel for promotion.....................	X	X	X			X		X		
67. Operate portable generators....	X								X	
68. Conceal activity/location with smoke grenades................	X	X			X					

Continued

TESTING PROBLEMS IDENTIFIED FOR A SAMPLE
OF 100 RECONNAISSANCE SPECIALIST JOB TASKS

TASK	TASK CONDITIONS									TASK BEHAVIOR		
	SCARCE	Equipment	Terrain	Personnel	DANGEROUS	VARIABLE	Surround	Display	LATENT	LONG PROCESS	TRANSIENT PROCESS	AFFECTIVE
69. Receipt ammunition............		X			X							
70. Emplace/recover expedient early warning devices.........			X					X				
71. Plan/coordinate employment of expedient early warning devices .			X	X				X				X
72. Avoid contact with poisonous plants.......................			X		X			X	X			
73. Locate a point on the ground using the polar coordinate system......................			X									
74. Navigate from one point on the ground to another with the aid of a compass.................			X							X	X	
75. Determine sequence for applying first aid measures to a casualty.				X	X			X				
76. Apply first aid measures for fracture, sprain, or dislocation				X	X			X			X	
77. Apply first aid measures for trench foot and immersion foot..				X				X			X	
78. Apply first aid measures for minor wounds.................				X				X			X	
79. Apply psychological first aid measures....................				X	X			X			X	
80. Construct field sanitation facilities....................		X								X		
81. Escape from burning tracked vehicle......................		X			X			X	X		X	
82. Load/unload an M151 series vehicle according to a loading plan		X										
83. Prepare a tracked vehicle for water operations (M113A1 or M577A1).....................		X		X								
84. Zero tracked vehicle gun launcher M551.................		X	X	X							X	
85. Apply immediate action in the case of the M551 failure to fire		X	X	X	X				X		X	
86. Operate an M151 series vehicle under reduced visibility.......		X	X				X				X	

Continued

TESTING PROBLEMS IDENTIFIED FOR A SAMPLE
OF 100 RECONNAISSANCE SPECIALIST JOB TASKS

TASK	TASK CONDITIONS								TASK BEHAVIOR		
	SCARCE	Equipment	Terrain	Personnel	DANGEROUS	VARIABLE	Surround Display	LATENT	LONG PROCESS	TRANSIENT PROCESS	AFFECTIVE
87. Guide a wheeled vehicle.......		X	X	X						X	
88. Mount/attach accessories on an M151 series vehicle........		X									
89. Adjust fire of M139 on a moving target................		X	X	X	X					X	
90. Adjust fire of M139 on a stationary target using BOT...		X	X	X	X					X	
91. Destroy tank and recon vehicle mounted machineguns by demolition..................		X	X		X					X	
92. Identify and correct malfunctions in a caliber .50 HBMT machinegun..................		X	X	X	X		X			X	
93. Fire caliber .50 machinegun from moving vehicle..........		X	X	X	X						
94. Perform checks and services on .50 caliber machinegun tripod and ammo.....................		X					X				X
95. Engage a moving target with an M60 machinegun............		X	X	X	X						
96. Apply immediate action to reduce stoppage on an M60 machinegun.........................		X	X	X	X			X		X	
97. Adjust headspace on an M2HBTT machinegun..................		X								X	
98. Perform maintenance services on caliber .45 pistol and magazine.....................							X				X
99. Identify cause of a caliber .45 pistol stoppage..........			X	X	X		X	X		X	
100. Perform preventative maintenance on canvas items.......							X				

APPENDIX B

PERFORMANCE AND KNOWLEDGE TESTS FOR THE TASK OF INSTALLING THE FIELD TELEPHONE

PERFORMANCE TEST

PERFORMANCE TEST
Put Telephone Set TA-312/PT Into Operation

TEST ORIENTATION
(Tester will read to Trainee)

"For this test you will put the field telephone into operation. The telephone is to be operated at the company level. One end of the electrical wire will be connected to switchboard SB-993/GT. The power source will be the 2 batteries. The task is to set up the telephone and contact another station. There is no time limit but work as quickly as you can. Do you have any questions?

NECESSARY EQUIPMENT

1 TA-312/PT telephone
2 BA-30 batteries
5" WD-1 field wire
1 pair pliers
1 screwdriver

TEST CONDITIONS

Telephone set will be in carrying case without batteries. Remianing equipment will be displayed on a table. Selector switch will be set at CB, INT-EXT switch will be set at EXT, and Volume Control will be set mid-way between LOW and LOUD. Test will be conducted indoors.

PERFORMANCE MEASURE 1	GO	NO GO
Set Up Field Telephone		
a. Strip approximately 1" from end of field wire	___	___
b. Insert bare wire into binding post	___	___
c. Install batteries -- 1 with electrode up, the other with electrode down	___	___
d. Set Circuit Selector Switch to LB	___	___
e. Set INT-EXT Switch to INT	___	___
f. Set Volume Control Knob to LOUD	___	___

PERFORMANCE MEASURE 2	GO	NO GO
Signal Station		
a. Hold receiver securely in retaining cradle	___	___
b. Turn generator crank	___	___

NOTE - If soldier forgets to signal station, remind him that is part of the test

WRITTEN CHOICE

Instructions

This is a regular multiple choice test. Choose the best answer for each question and write the letter by the correct answer on your answer sheet.

The telephone is to be operated at the company level with 2 batteries as the power source.

These 7 questions relate to putting the Telephone Set, TA-312/PT into operation. For these questions, assume that the telephone is being operated at the company level connected to switchboard SB-993/GT. Power source is 2 BA-30 batteries.

1. How should field wire be secured to the TA-312/PT?

   a. Connected to binding posts.
   b. Connected to external battery terminals.
   c. Connected to external terminals.

2. How should the two batteries in the TA-312/PT be installed?

   a. One with positive (electrode) end up, the other with negative (smooth) end up.
   b. Both with positive (electrode) end up.
   c. Both with negative (smooth) end up.

3. How should the field wire be prepared?

   a. Last three inches should be stripped.
   b. Last inch should be stripped.
   c. Insulation should be left intact.

4. What is the correct position for the circuit selector switch when putting the TA-312/PT into operation at the company level?

   a. CBS
   b. CB
   c. LB

5. What is the correct position for buzzer volume control knob when putting the TA-312/PT into operation at the company level?

   a. LOW
   b. Half way between LOW and LOUD
   c. LOUD

6. What position should the EXT-INT switch be in when the handset is used with the TA-312/PT?

   a. EXT
   b. INT

7. How does operator signal a station with the TA-312/PT?

   a. Press push-to-talk switch.
   b. Turn generator crank with handset securely in retaining cradle.
   c. Turn generator crank with handset removed from retainer cradle.

PICTURE CHOICE

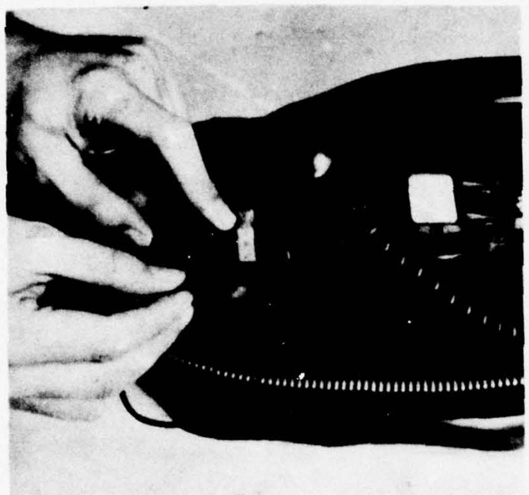
Instructions

This test has 7 questions. The questions are in words like most tests but the possible answers are pictures. Choose the picture that best answers the question and write the letter by that picture on your answer sheet.

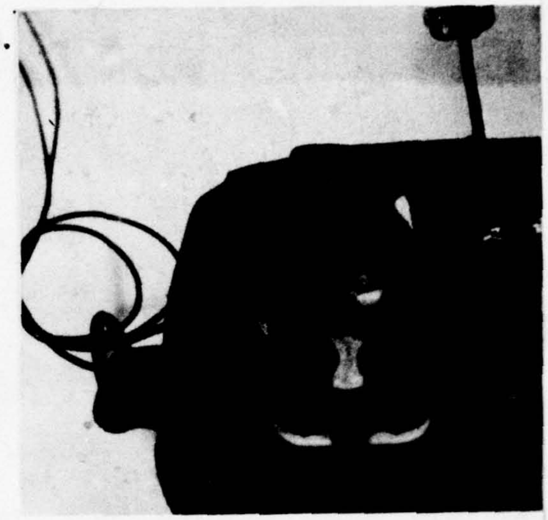
Again, the telephone is to be operated at the company level with batteries as the power source.

1. How should field wire be secured to the TA-312/PT?

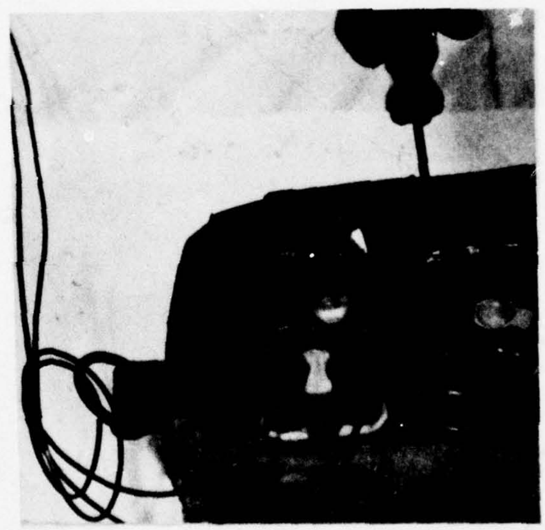
a.



b.

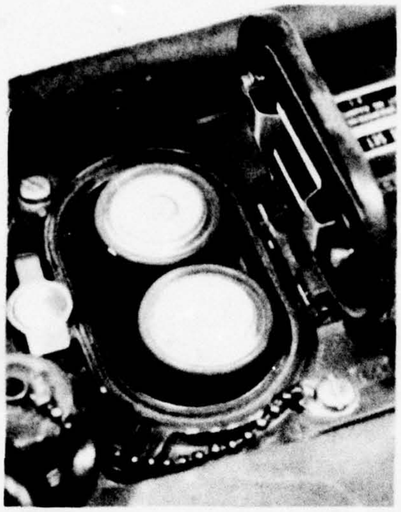


c.

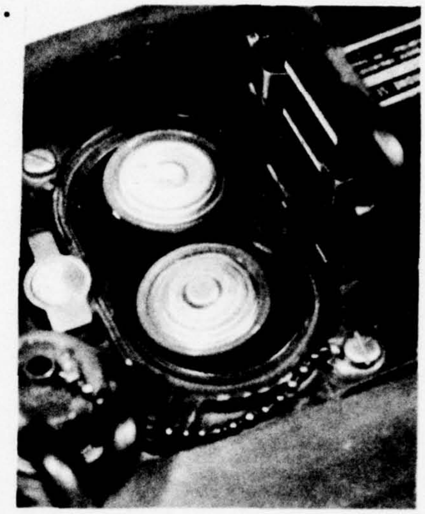


2. How should the two batteries in the TA-312PT be installed?

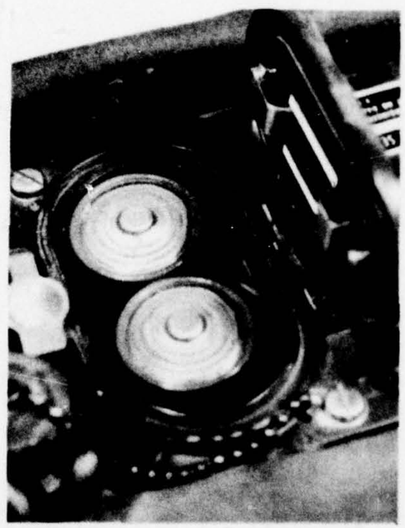
a.



b.

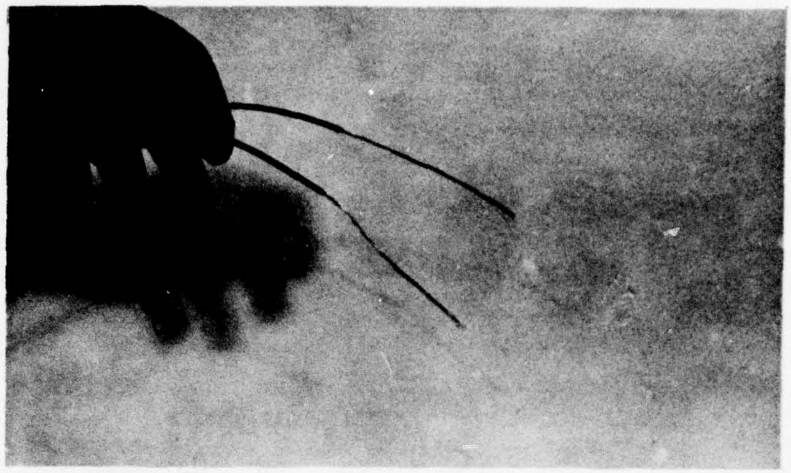


c.

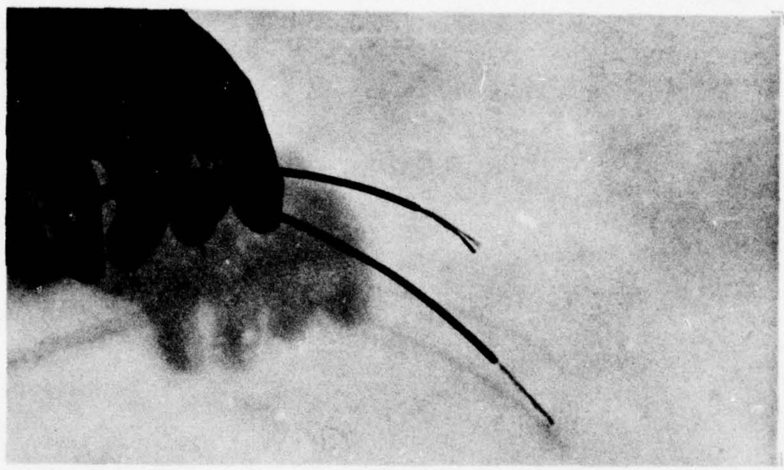


3. How should the field wire be prepared?

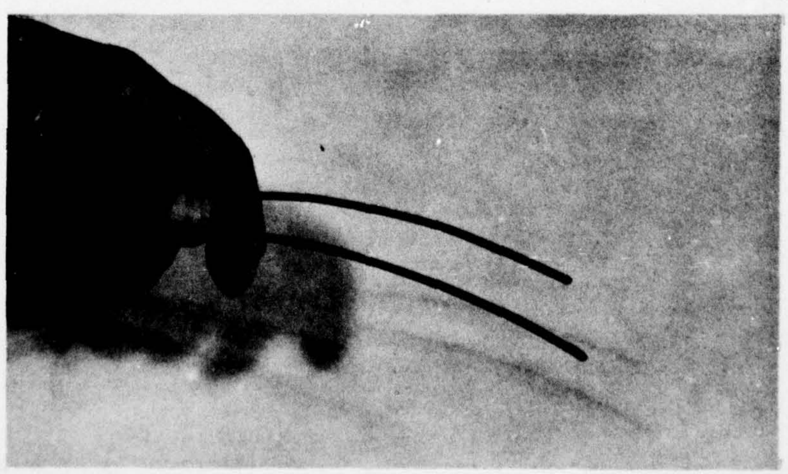
a.



b.



c.



4. What is the correct position for the circuit selector switch when putting the TA-312/PT into operation at the company level?

a.

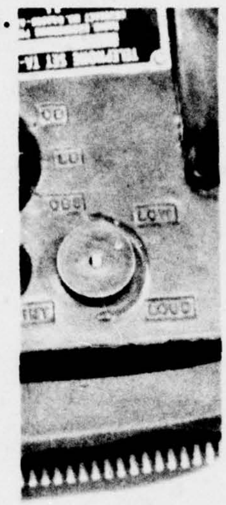
b.

c.

5. What is the correct position for buzzer volume control knob when putting the TA-312/PT into operation at the company level?

a.

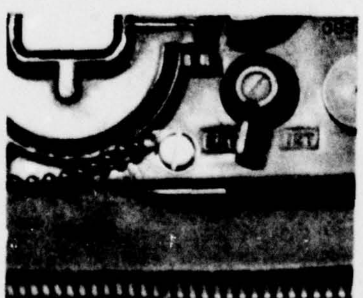
b.



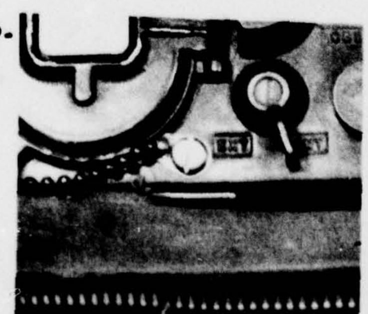
c.

6. What position should the EXT-INT switch be in when the handset is used with the TA-312/PT?

a.

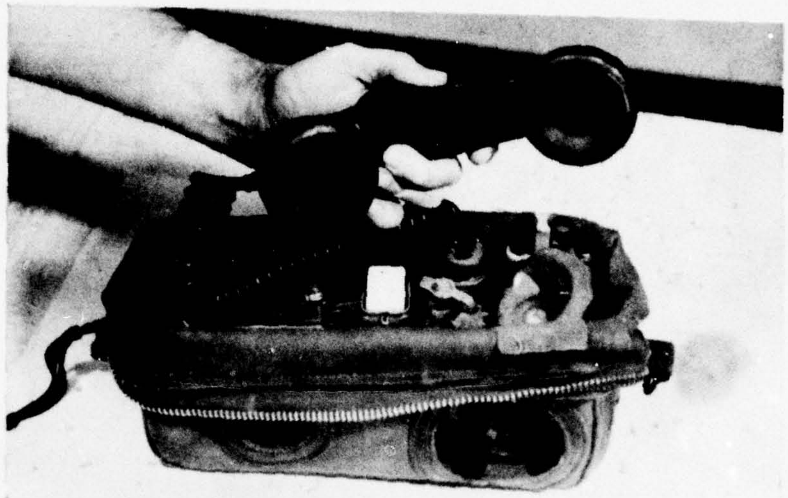


b.

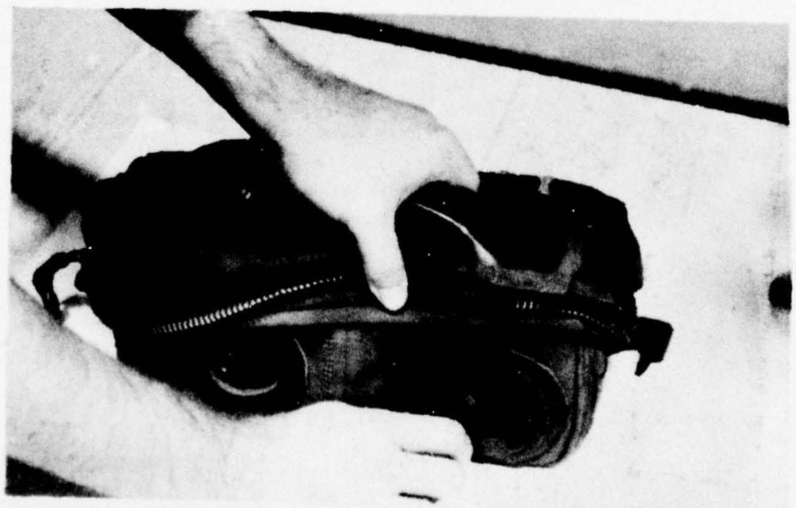


7. How does operator signal a station with the TA-312/PT?

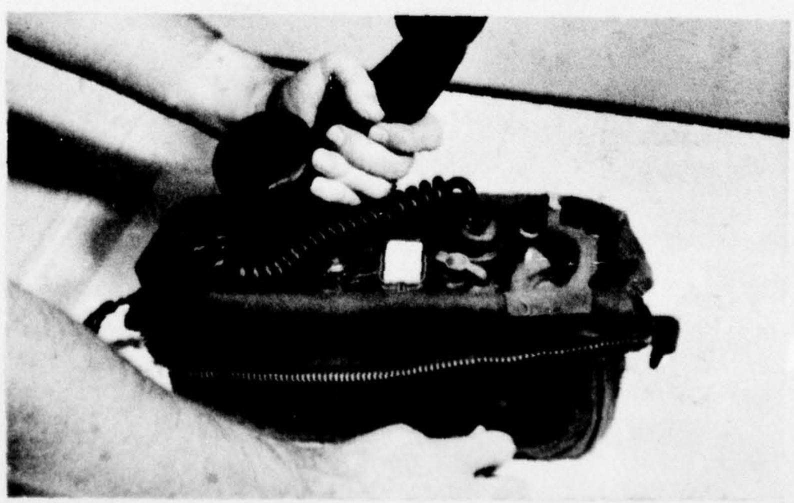
a.



b.



c.



PICTURE OUTCOME

Instructions

This is a picture of a soldier contacting another station after setting up the field telephone. Study the picture. If there are any errors in the way he set up the telephone or the way he is signaling the station, circle the error with your grease pencil.


TELEPHONE SET TA 312 PT
SERIAL NO

PICTURE SORT

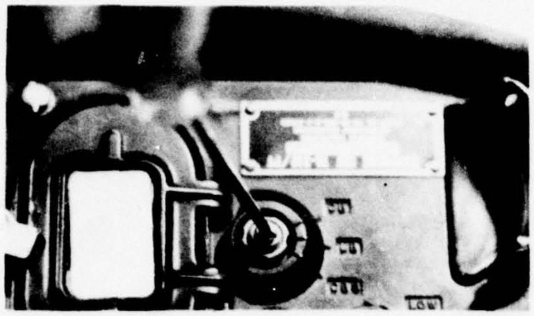
Instructions [1]

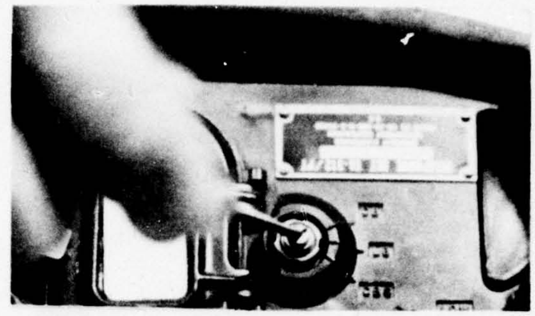
Here are some pictures of a soldier setting up a field telephone and contacting another station. Choose the pictures that show how you would do each step and put them in the correct order. You do not need to use every picture.

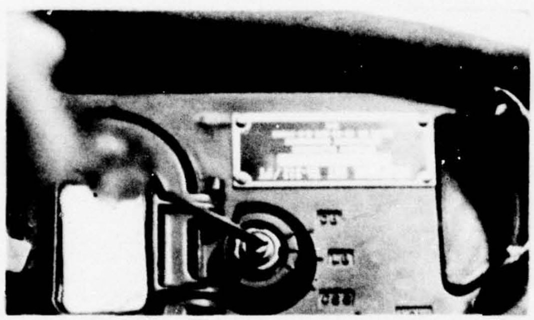
There is no time limit, but work as quickly as you can. Remember the task is to set up the field telephone at the company level and contact another station.

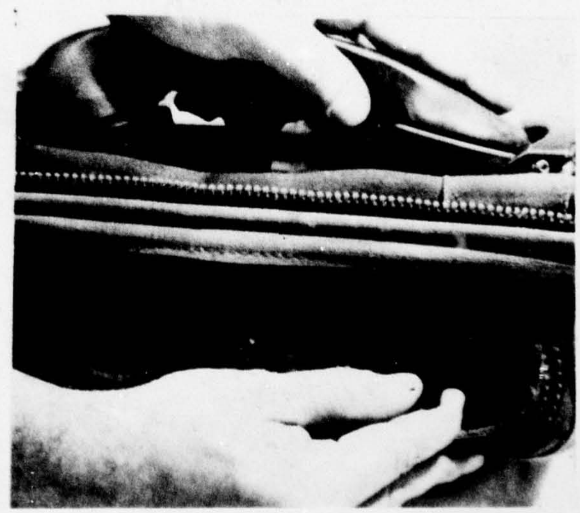
---

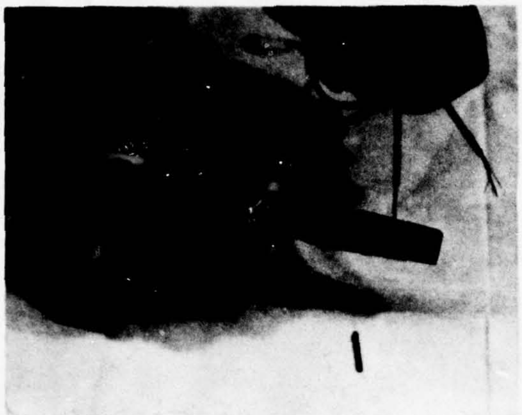
[1] The photographs were pasted individually on 5" x 9" cards and stacked in the order shown for presentation to S.







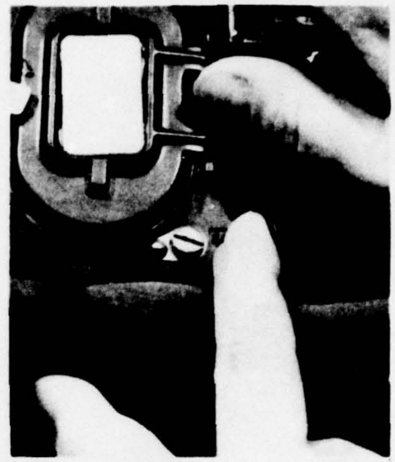


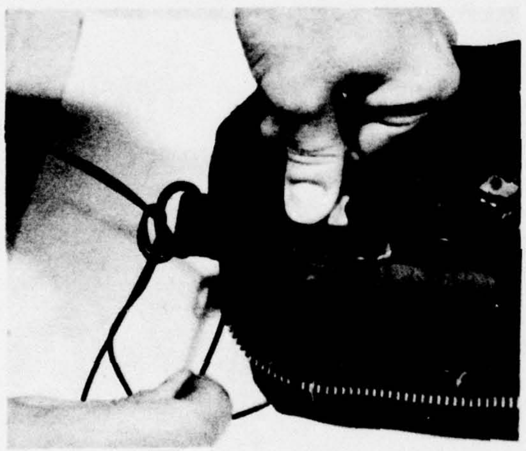


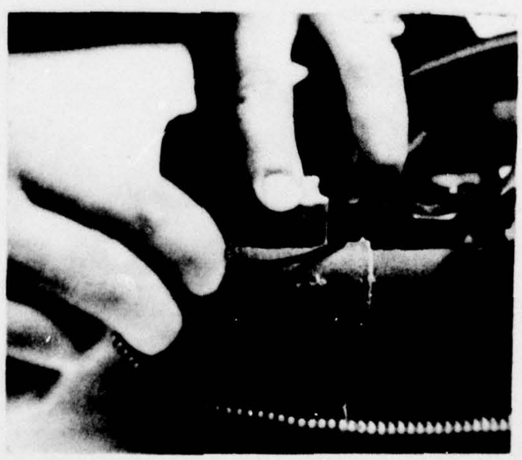
AD-A060 338 HUMAN RESOURCES RESEARCH ORGANIZATION ALEXANDRIA VA F/G 5/10
RESEARCH ON METHODS OF SYNTHETIC PERFORMANCE TESTING.(U)
APR 76 W C OSBORN, J P FORD DAHC19-74-C-0059
UNCLASSIFIED HUMRRO-FR-CD(L)-76-1 NL

2 OF 2
ADA
060338

END
DATE
FILMED
12-78
DDC





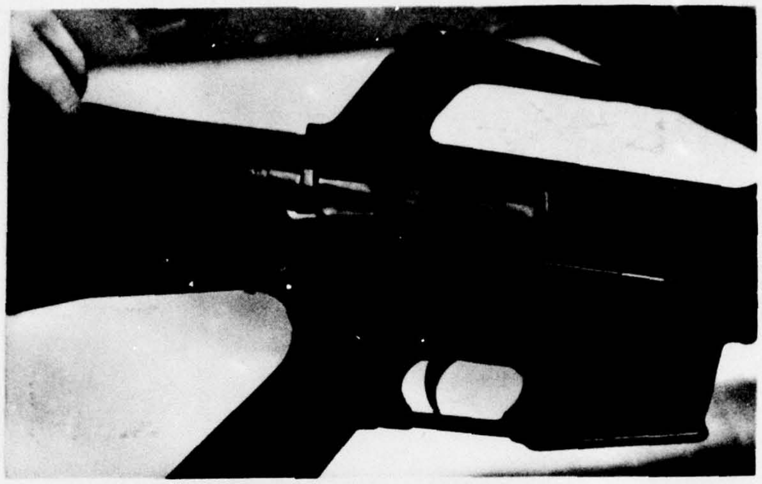


APPENDIX C

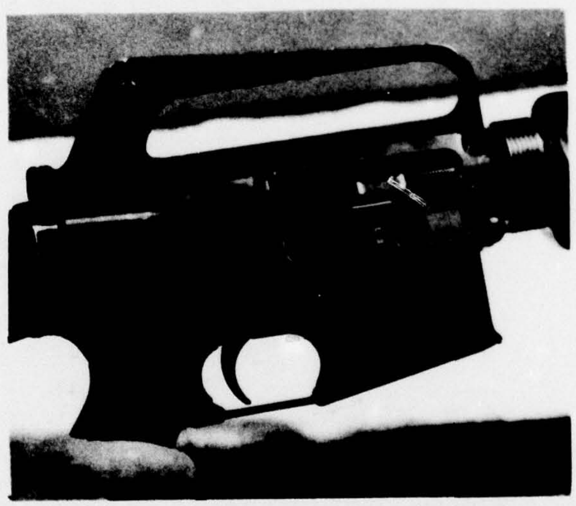
SAMPLE PC QUESTION ILLUSTRATING SUB-STEP SEQUENCES AS ANSWER ALTERNATIVES

1. After inspecting the receiver and chamber for ammunition, the step(s) *to complete clearing the M16 rifle for disassembly are:*

a.



b.



c.

APPENDIX D

SAMPLE PO QUESTION ILLUSTRATING USE OF BLOW-UPS TO PROVIDE DETAIL OF SPATIALLY DIVERSE COMPONENTS